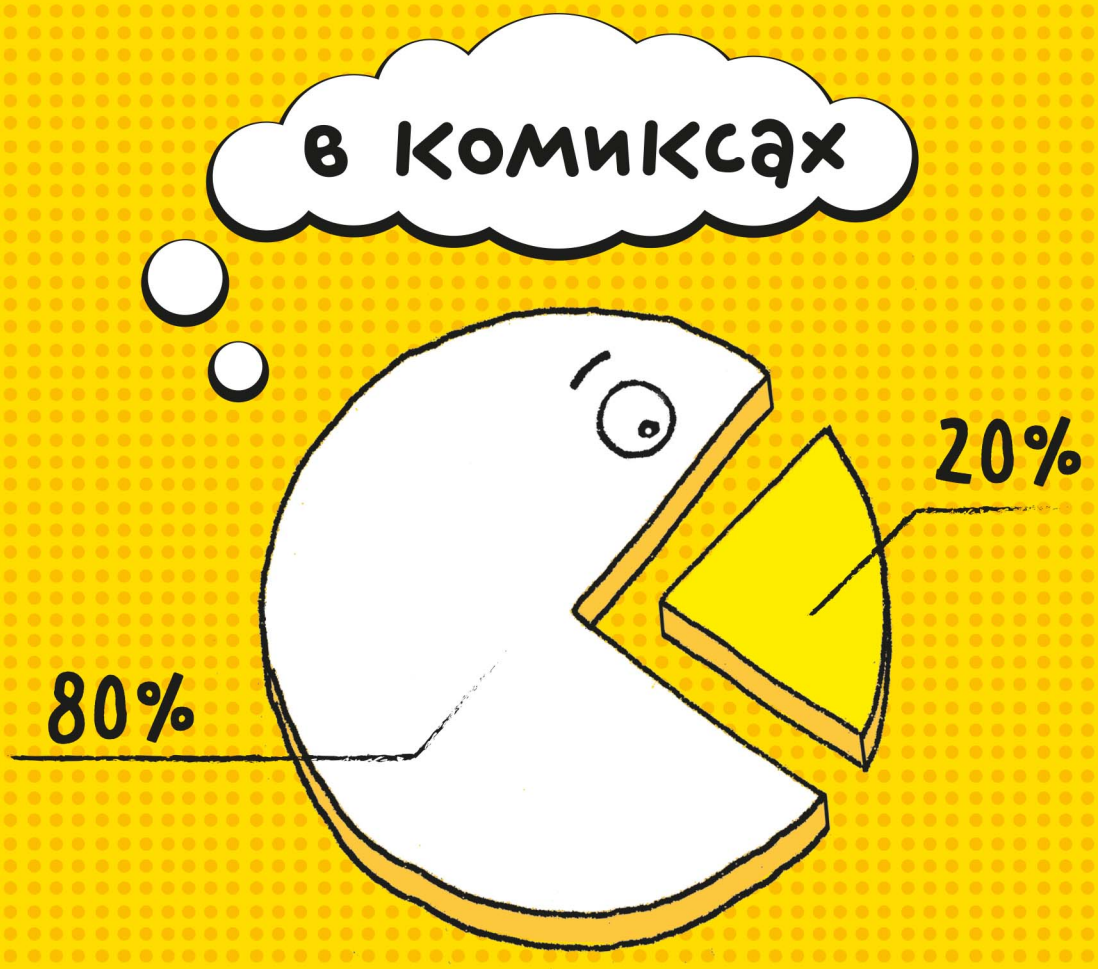
СТАТИСТИКА

в комиксах



Айлин Магнелло

доктор наук оксфордского университета

Берин Ван Лоон

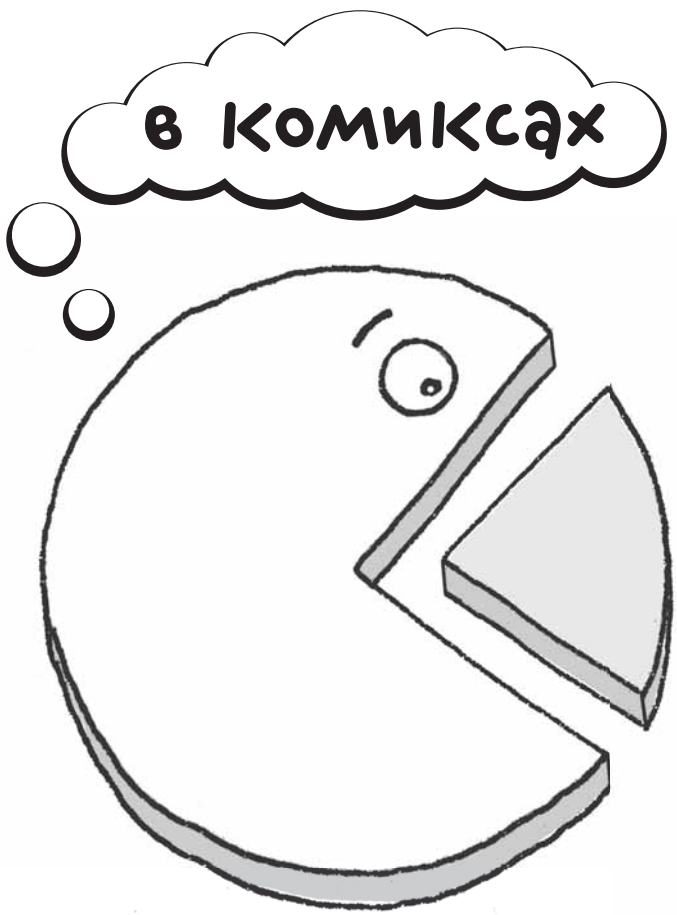
художник-сюрреалист, иллюстратор

Айлин Магнелло
доктор наук оксфордского университета

Берин Ван Лоон
художник-сюрреалист, иллюстратор

СТАТИСТИКА

в комиксах



БОМБОРА™
Москва 2018

УДК 311
ББК 60.6
М12

Introducing Statistics: A Graphic Guide
by Eileen Magnello, Borin Van Loon

Магнелло, Эйлин.

М12 Статистика в комиксах / Эйлин Магнелло, худож. Борин Ван Лоон ; [пер. с англ. Д. Кудряшова]. — Москва : Эксмо, 2018. — 176 с. : ил. — (Бизнес в комиксах).

ISBN 978-5-04-090149-4

Демографическая статистика против математической, вероятности, выборки, популяции, «жизненная статистика» Уильяма Фарра и математическая Карла Пирсона... — в этом комиксе обзор истории, философии, основные концепции и то, как они связаны с реальными проблемами. Решения, основанные на статистике, принимаются каждый день и влияют на нашу повседневную жизнь. От тестов на профпригодность, которые дают нам работодатели, одежды, которую мы носим, до еды, которую мы едим, и даже пива, которое мы пьем. Знание основ статистики может даже спасти или продлить жизни!

УДК 311
ББК 60.6

ISBN 978-5-04-090149-4

Погружаясь в числа

Мы погружаемся в статистику, и она состоит не только из чисел. У СМИ статистика вызывает страх и ужас, а иногда воодушевление. В печати авторы постоянно говорят о том, что статистика преступлений, болезней, бедности и задержек транспорта не является источником проблемы, а представляет реальных людей или субъектов, являясь чем-то большим, чем отметкой на графике.

Средние или вариативные значения?

Большая часть шокирующей статистической информации, которая распространяется через СМИ, основана на **средних значениях**. Несмотря на зачастую обманчивую озабоченность средними значениями, самая важная часть статистического концепта бывает опущена журналистами и репортерами, и эта часть — **вариативные значения**. Идея вариативности лежит в основе современной математической статистики и играет главную роль в биологической, медицинской, образовательной и промышленной статистике.

Почему же вариации так важны?

Вариации можно с легкостью наблюдать в мультикультурной Британии, в особенности в Лондоне, который сейчас состоит из более чем 300 культур, говорящих на многих языках (от ачоли до зулусского языка) и тринадцати различных верований. Для некоторых мультикультурализм состоит в ценности каждого индивида и сохранении уникальной культуры каждого индивида (а также в избегании сводить этнически различные группы индивидов к какому-то одному представителю).

Эти разнообразные индивидуальные различия заключают в себе идею статистической изменчивости, которая является основой современной математической статистики.

Зачем изучать статистику?

Статистика используется учеными, экономистами, чиновниками и промышленниками. Решения, основанные на статистике, принимаются каждый день и влияют на нашу повседневную жизнь — от лекарств, которые мы принимаем, лечебной помощи, которую нам оказывают, тестов на профпригодность, которые предлагают нам работодатели, машин, которые мы водим, одежды, которую мы носим (производители шерсти используют статистические тесты для определения нитей, которые будут максимально удобны), до еды, которую мы едим, и даже пива, которое мы пьем.

Знание основ статистики может даже спасать или продлевать жизни, как это случилось со Стивеном Джеем Гулдом (Gould), о котором мы расскажем чуть позже.

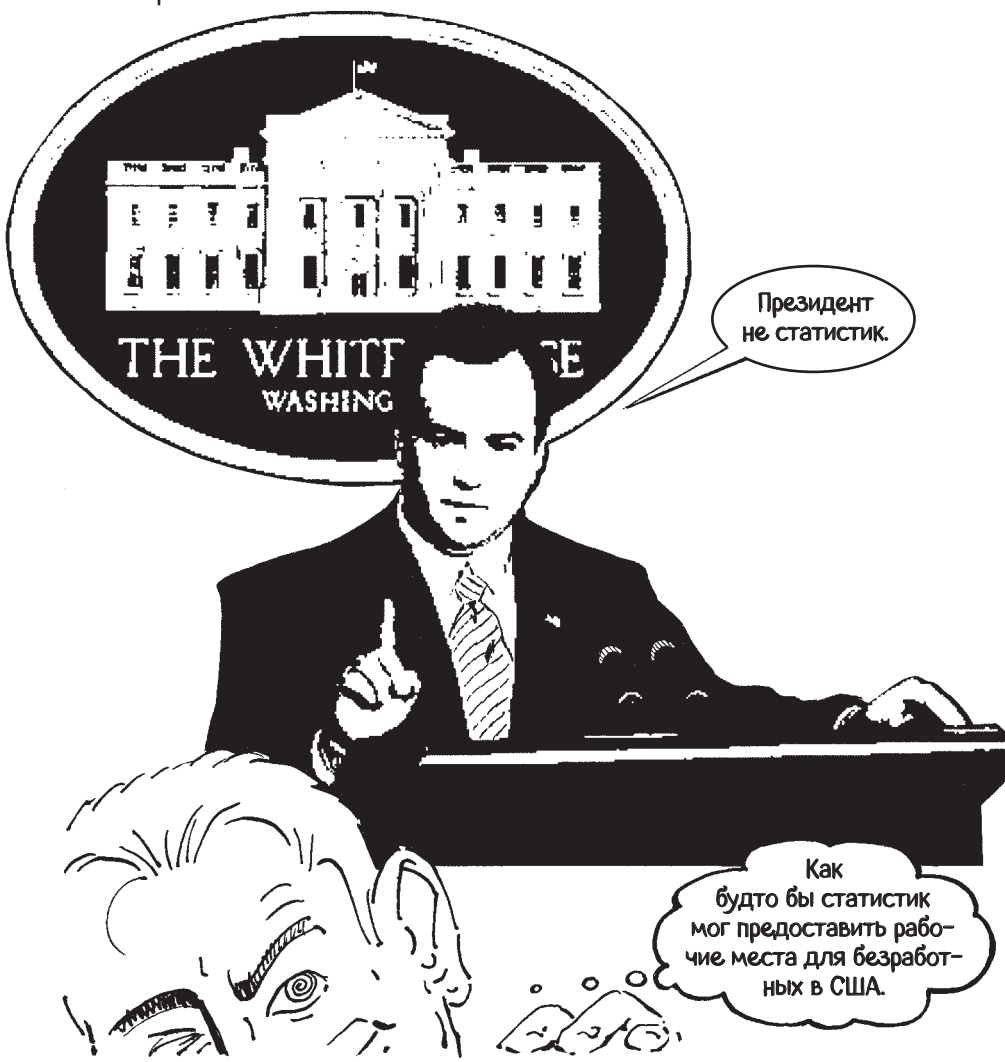
Статистика – что это?

Несмотря на широкое распространение, определить границы статистики очень сложно. Как сказал один колумнист: «Сигареты являются основной причиной статистики». Люди предпочитают избегать неприятных вещей, говоря: «Я не хочу быть очередной частью статистики». Но неужели ученые, занимающиеся статистикой, действительно полагают, что все человечество можно свести к нескольким числам?

Несмотря на то, что некоторые люди думают, что результаты статистики неопровержимы, другие верят, что они обманчивы.

Марк Твен приписал этот афоризм премьер-министру Великобритании Бенджамину Дизраэли в 1904 году. На самом деле Леонард Генри Кортни (Courtney) впервые сказал это в своей речи в Саратога-Спрингс, в Нью-Йорке в 1895 году, имея в виду пропорциональность представителей из 44 американских штатов.

Некоторые государственные чиновники даже обвиняют статистику в создании экономических проблем. Когда пресс-секретарь Белого дома Скотт Макклеллан (McClellan) в феврале 2004 года попытался объяснить, почему администрация президента Буша отказалась от своего прогноза, который предсказывал увеличение количества рабочих мест в США, его объяснение было простым.



В Великобритании Комитет по статистике призывал, чтобы «Членам Кабинета министров было запрещено проверять статистические данные до их публичной огласки, так как это поможет избежать политического давления или эксплуатации». Тем не менее статистика, которая доступна в публичном поле, может формировать мнение граждан, влиять на государственную политику и информировать (или дезинформировать) граждан о медицинских и научных открытиях и прорывах.

Что означает слово «статистика»?

Слово «статистика» произошло от латинского *«status»,* которое в свою очередь перешло в итальянский как *«statista»* и впервые было использовано в XVI веке, обозначая государственников или государственных деятелей — тех, кто был связан с делами государства. Немцы стали использовать слово Statistik около 1750 года, французы ввели слово statistique в 1785 году, а голландцы создали термин statistiek в 1807 году.

Система статистики была впервые использована в XVII веке английским купцом **Джоном Граунтом** (1620–1674) и ирландским естествоиспытателем и экономистом **Уильямом Петти** (1623–1687).

В XVIII веке многие ученые-статистики были юристами, их образование было в сфере общего права (ветвь права, занимающаяся государством).

Шотландский землевладелец и первый президент министерства сельского хозяйства сэр **Джон Синклер** (1754–1834) был первым, кто ввел термин *«статистика»* в английский язык в 1798 году в своей работе «Статистический отчет о Шотландии».

Синклер использовал статистику для анализа общественных явлений вместо политических. Это привело к развитию демографической статистики в середине XIX века.

Демографическая статистика vs математическая статистика

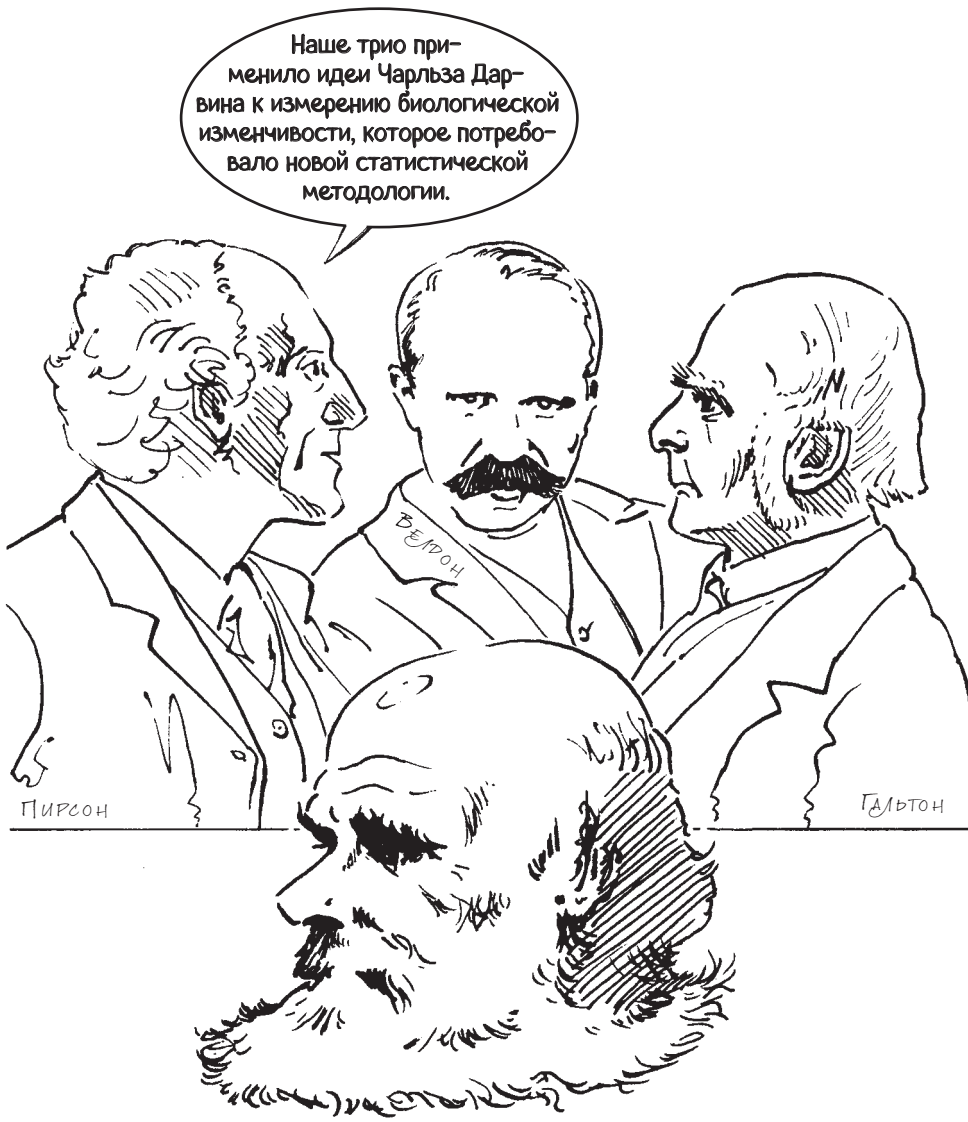
Не вся статистика одинакова. Есть два типа: демографическая и математическая статистика.

Демографическая статистика — это та, которую бóльшая часть людей понимает под статистикой. Она является совокупностью собранных данных.

Этот процесс связан в основном со средними значениями и использует таблицы продолжительности жизни, проценты, пропорции и коэффициенты: вероятность используется большей частью в актуарных (т. е. при страховании жизни) целях. Только с начала XX века слово *«статистика»* стало использоваться для обозначения отдельного факта.

Математическая статистика появилась как ветвь математической теории вероятностей в конце XVIII века в работах таких континентальных математиков, как Якоб Бернулли, Абрахам де Муавр, Пьер-Симон Лаплас и Карл Фридрих Гаусс.

В конце XIX века математическая статистика начала оформляться в полноценную науку благодаря работам **Фрэнсиса Исидро Эджуорта** (1845–1926), **Джона Венна** (1834–1923), **Фрэнсиса Гальтона** (1822–1911), **Уолтера Фрэнка Рафаэля Велдона** (Weldon) (1860–1906) и **Карла Пирсона** (1857–1836).



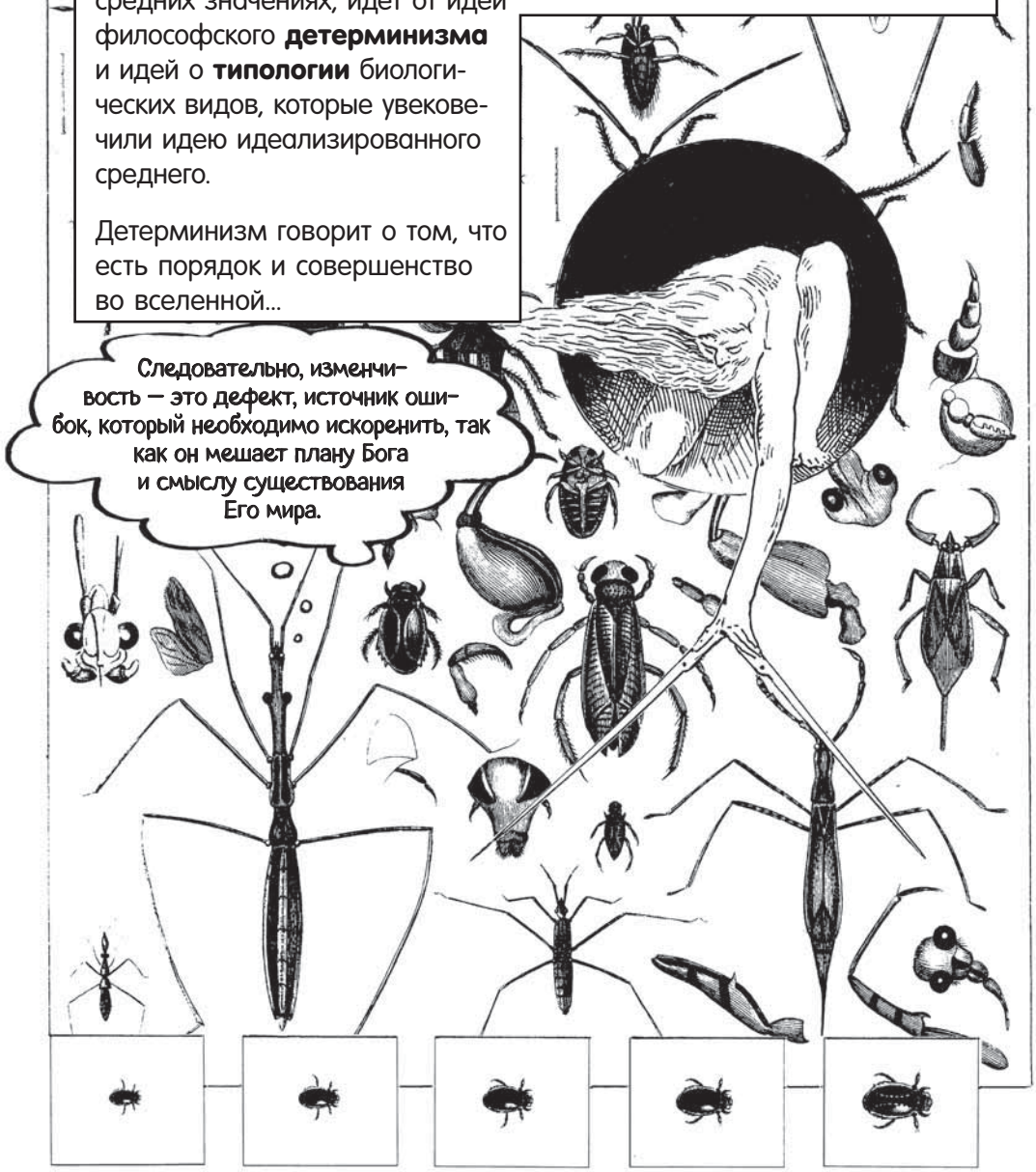
Математическая статистика охватывает научные дисциплины, включая анализ изменчивости, в основе которого лежит матричная алгебра. Математическая статистика имеет дело со сбором, классификацией, описанием и интерпретацией данных, полученных из соцопросов, научных экспериментов и клинических испытаний. Вероятность используется для установления критериев статистической значимости и соответствующих статистических тестов.

Если говорить в таком ключе, статистика является технической наукой, и коль скоро речь идет о математической статистике, необходимо понимать статистические идеи, которые лежат в основе математических методов.

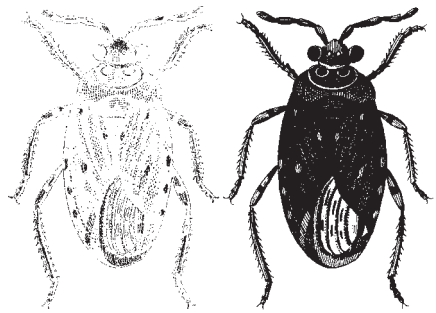
Философия статистики

Выбор изучения средних значений или же измерения вариативных значений (изменчивости) уходит корнями в философские идеи, над которыми размышляли ученые-статистики, естествоиспытатели и математики на протяжении XIX века. Акцент, сделанный на статистических средних значениях, идет от идеи философского **детерминизма** и идей о **типологии** биологических видов, которые увековечили идею идеализированного среднего.

Детерминизм говорит о том, что есть порядок и совершенство во вселенной...

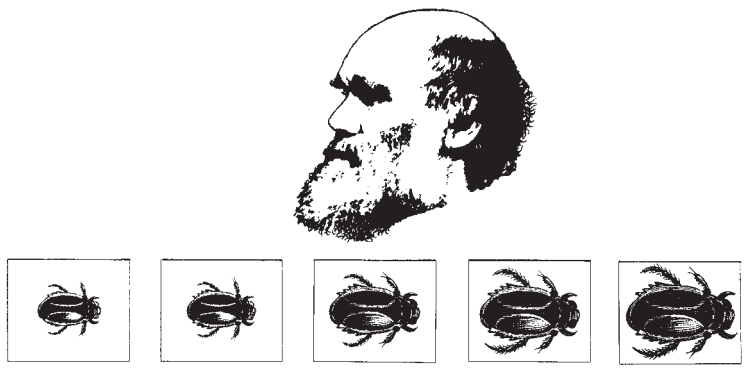


Типологическое рассмотрение *видов*, которое властвовало над умами таксономистов*, типологов и морфологов до конца XIX века, способствовало росту популярности морфологического понятия *вида*. Согласно этому понятию, виды рассматривались как представители идеального типа.



Существование идеального типа было выведено из некоторого морфологического сходства, по критерию которого типологи различали виды. Из этого наблюдения можно было сделать вывод, что количество видов быстро растет, так как любое отклонение от типа приводит к классификации новых видов.

Подлинное изменение, согласно морфологическому понятию вида, возможно только благодаря **скачкообразным** образованиям новых видов, имея в виду, что новые виды возникают скачкообразно в каждом отдельном поколении. Так как теория эволюции Дарвина говорит о **«постепенных»** изменениях, она была несовместима с эссенциализмом**.



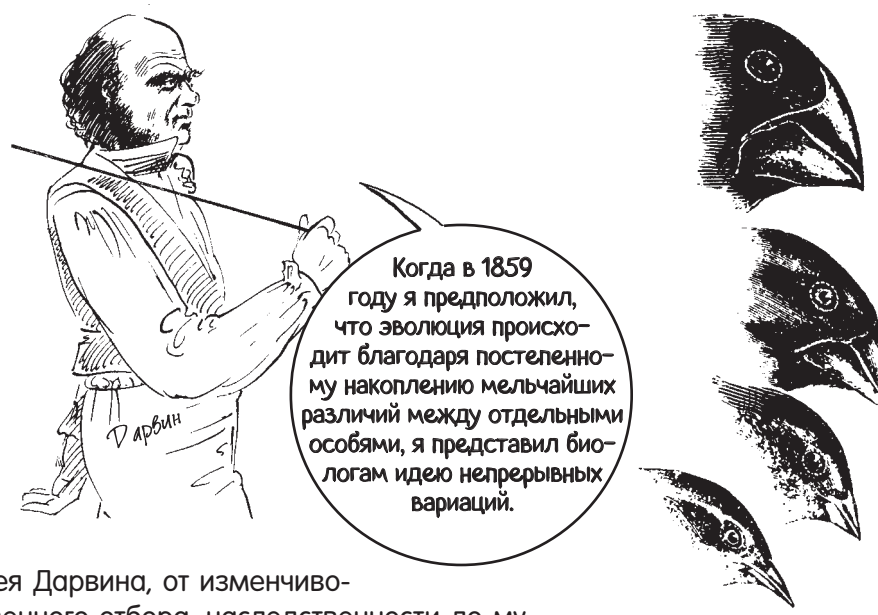
---

* Таксономисты классифицируют организмы по группам.
Типологи классифицируют организмы согласно общим типам.
Морфологи изучают строение организмов.

**Эссенциализм *(от лат. essentia — «сущность»)*, — философское учение, согласно которому у каждой вещи есть глубинная скрытая реальность, истинная природа, характеризующаяся неизменным набором качеств и свойств. Возникло и развивалось в Средние века в рамках схоластической философии. — *Прим. науч. ред.*

Дарвин и статистические популяции

Переход к измерению статистической изменчивости хорошо просматривается в идеологическом сдвиге, который произошел в середине XIX века, когда **Чарлз Дарвин** (1809–1882) начал изучать мельчайшие биологические вариации у растений и животных.



Каждая идея Дарвина, от изменчивости, естественного отбора, наследственности до мутаций к исходному виду*, требовала статистического анализа.

Дарвин не только показал, что изменчивость можно измерять и получать ценную информацию, рассматривая статистические популяции, а не отдельные типы или сущности, но он также занимался различными типами взаимосвязи (correlation), которые можно было бы использовать для объяснения естественного отбора.

Как сказал в 1931 году биолог-эволюционист **Сьюалл Райт** (Wright) (1899–1988):

---

* В генетике — так называемая реверсная мутация, т. е. восстановление у мутантного организма исходной структуры ДНК. — *Прим. науч. ред.*

Викторианские ценности

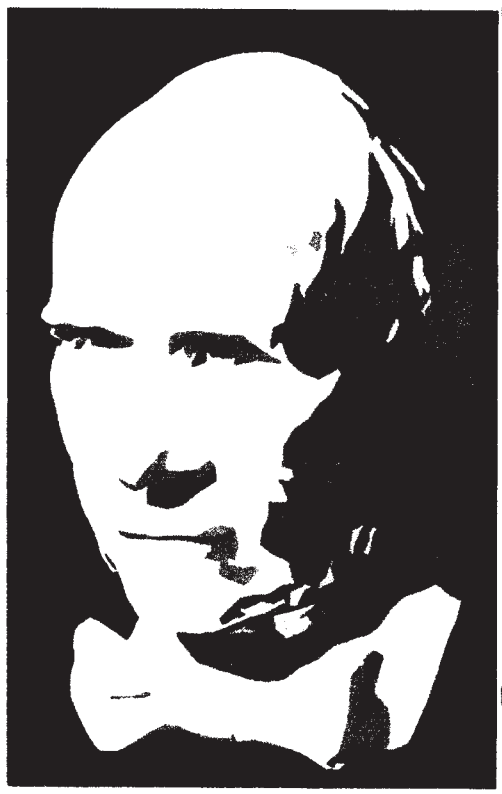
Несмотря на определенное развитие демографической и математической статистики в континентальной Европе, стремительному росту демографической статистики в середине XIX века и математической статистики на рубеже XIX–XX веков мы обязаны этим викторианцам*.

Развитие обеих ветвей статистики происходило в широком контексте викторианской культуры измерений. Викторианцы высоко ставили точность и аккуратность, как в материальных, так и в духовных сферах, потому что это позволяло получать более надежную информацию. В расширяющейся индустриальной экономике было необходимо получить те результаты, которые затем можно было бы повторно воспроизвести на международном рынке.

---

* В нижеследующем перечне ученых явным образом не хватает крупных немецких (В. Лексис) и российских (В. И. Борткевич, А. А. Чупров, Е. Е. Слуцкий) исследователей. — *Прим. науч. ред.*

Инженеры и физики днями и ночами работали в лабораториях, записывая и измеряя электрические, механические и физические постоянные для машин, оборудования и прочих объектов. Биологи и геологи собирали как можно больше данных в своих экспедициях для создания географических карт, измеряли долготу и широту и классифицировали новые виды растений и животных.



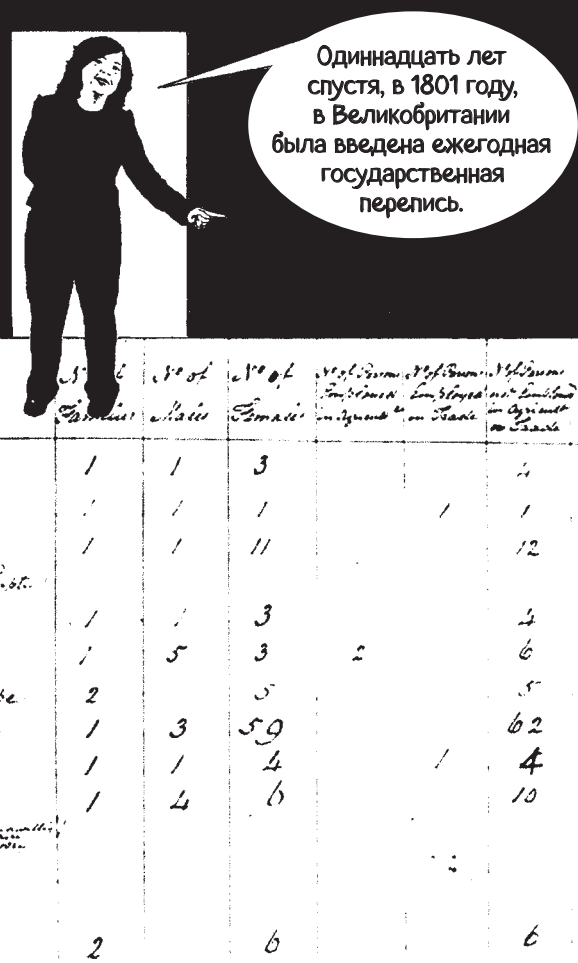
Доктор Джон Сноу

Статистика предложила способ, которым можно было определить количество измерений в сфере жизнедеятельности человека, в особенности тех, которые касались здоровья и гигиены граждан, эпидемий, наследственности и медицины.

С чего все началось?

Подсчет населения и проведение переписей является одной из наиболее древних известных человечеству практик статистики: вавилоняне, египтяне и китайцы собирали статистические данные о своем населении, в основном для поиска пригодных для военной службы граждан, а также для установления ставок налоговых сборов. В первом тысячелетии до нашей эры римляне и греки проводили переписи. Слово «Census» происходит от римских цензоров, чьей обязанностью было подсчитывать количество людей. Римская перепись состояла из списка граждан Рима и их собственности.

Скандинавские страны ввели первые национальные переписи в середине XVII века. Первая перепись, проведенная в США в 1790 году, показала пропорциональное представительство при выборе конгрессменов в тринадцати американских штатах.



Метрические книги

Как подсчитывали людей до введения государственных переписей? Метрические книги давали ценную информацию для понимания некоторых ранних идей о населении. В начале XIV века во Франции, в Бургундии, регистрировали смерти и браки, а к XVI веку регистрация крещений, браков и смертей стала обязательной для французских священников. В Англии ответственность за сбор подобной информации в 1538 году была возложена на местное духовенство благодаря Томасу Кромвелю, лорду-канцлеру короля Генриха VIII.

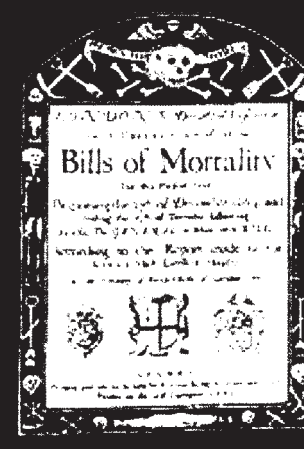
Однако раскольники и люди иных вероисповеданий были исключены из записей, так же как и многие другие внутри англиканской церкви, которые не желали или не могли позволить себе оплачивать регистрацию в церкви.

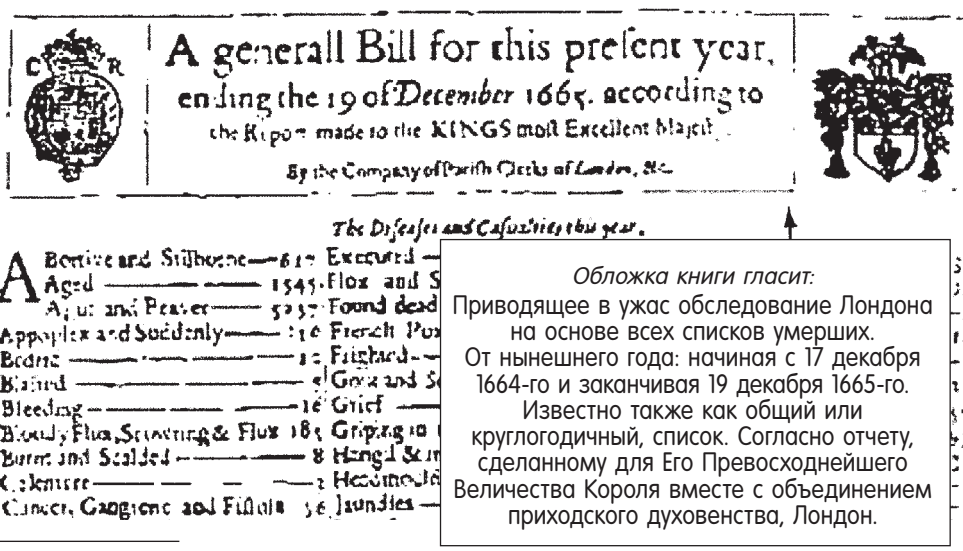
Лондонские списки умерших

На протяжении XVII и XVIII веков в Англии росло количество людей, которые придерживались религий, отличных от официальной. Несмотря на то что иудеям разрешалось собирать информацию, квакерам*, а также другим церквам, отделенным от государства, не позволялось собирать данные, так как они считались неприемлемым источником для государства, находясь вне официальной системы.

С учетом того что большое количество людей не разрешалось подсчитывать, нарастал интерес к тому, убывает или же растет население Англии.

Джон Граунт был одним из первых, кто в своей работе «Естественные и политические наблюдения на основе лондонских списков умерших» попытался использовать информацию 10 000 метрических записей Англии и Уэльса, в которую входили сведения о поле, возрасте и причине смерти. Граунт использовал термин «политическая арифметика» для описания своей работы — этот термин ему подсказал его друг Уильям Петти.




A generall Bill for this present year,
ending the 19 of December 1665. according to
the Report made to the KINGS most Excellent Majesty
By the Company of Parish Clerks of London, &c.

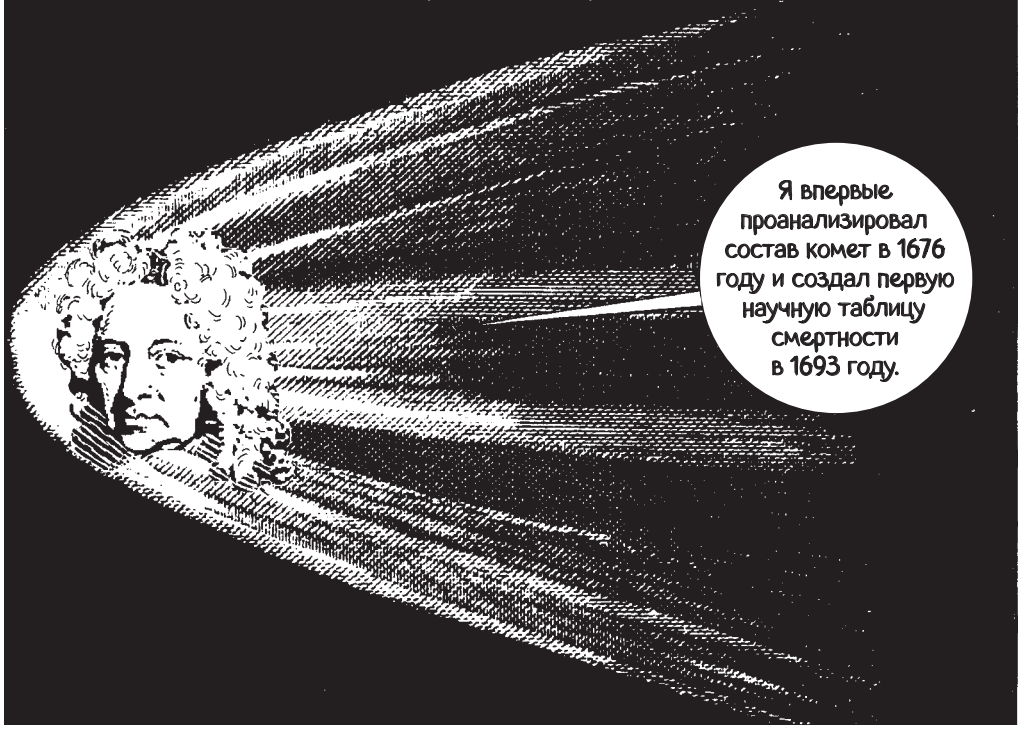
The Diseases and Casualties this year.

Обложка книги гласит:
Приводящее в ужас обследование Лондона на основе всех списков умерших. От нынешнего года: начиная с 17 декабря 1664-го и заканчивая 19 декабря 1665-го. Известно также как общий или круглогодичный, список. Согласно отчету, сделанному для Его Превосходнейшего Величества Короля вместе с объединением приходского духовенства, Лондон.

* Протестантское христианское движение, возникшее в годы революции в середине XVII века в Англии и Уэльсе и характеризующееся независимостью своих религиозных организаций и объединений. — *Прим. науч. ред.*

Таблицы смертности Галлея

Выдающийся труд, основанный на данных о смертности в XVIII веке, был связан с созданием таблицы продолжительности жизни. Идея была предложена Джоном Граунтом, а затем реализована **Эдмундом Галлеем** (1656–1742), имя которого было присвоено известной комете.



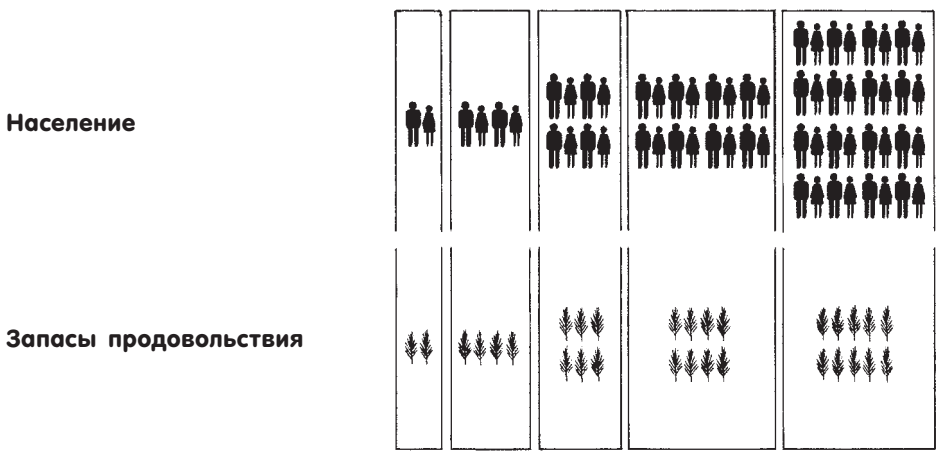
Голландский астроном и политический арифметик **Николас Стрюик** (Struyck) (1687–1769) создал свои работы на основе трудов Галлея о кометах и о размерах населения. Стрюик провел широкомасштабные исследования населения в Нидерландах, но его главной целью была оценка приблизительного количества населения Земли. Его волновал вопрос о том, растет, остается таким же или же сокращается население Земли*.

* Сравнительно недавно заново открытое имя в науке, в классическом исследовании И. Тодхантера (История математических теорий притяжения и фигуры Земли от Ньютона до Лапласа / Пер. с англ. М., 2002) сведений о Стрюике нет. — *Прим. науч. ред.*

Мальтузианское население

Пока многие ученые пытались подсчитать население страны или мира, экономист **Томас Роберт Мальтус** (1766–1834) в своей знаменитой работе «Опыт закона о народонаселении» (1798) показал, что неконтролируемое людьми население Земли будет постоянно увеличиваться (превышая необходимые средства существования) и что улучшение человеческой жизни может быть достигнуто жесткими ограничениями рождаемости.

Мальтус полагал, что население растет экспоненциально (2, 4, 8, 16, 32 и т. д.), в то время как запасы продовольствия растут по арифметической линейной прогрессии (2, 4, 6, 8, 10 и т. д.). Гипотеза Мальтуса говорит о том, что живущее в каждый настоящий момент население будет иметь тенденцию превышать запас имеющегося продовольствия.



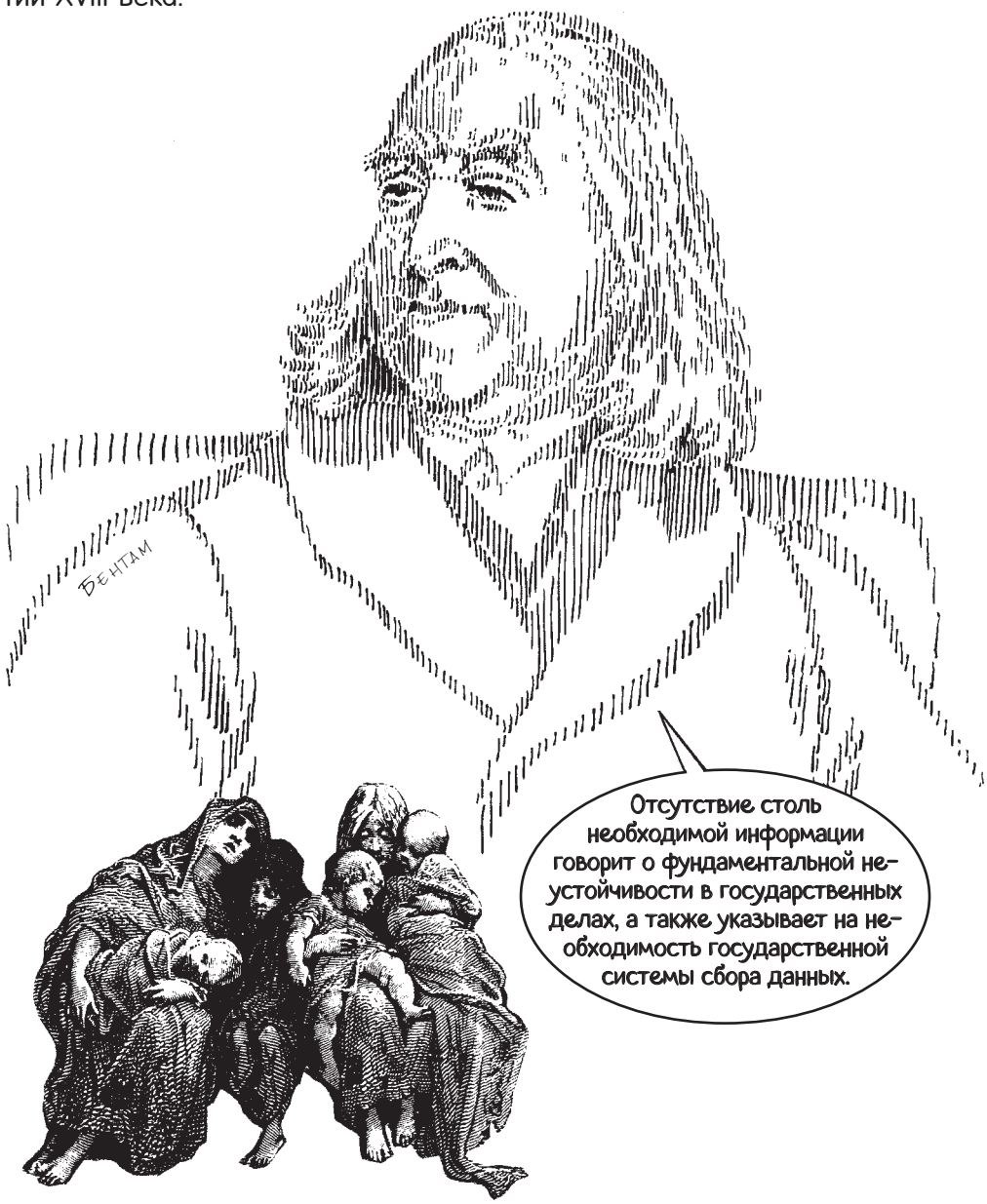
Демография — наука о населении

Всякая попытка улучшения условий жизни низших слоев населения путем увеличения их доходов или развития сельскохозяйственного производства казалась Мальтусу бесполезной. Он считал, что для сокращения роста населения необходима «сдерживающая сила морали». **Демография** началась как количественное изучение бедности.

В конце XVIII века Мальтус считал, что рост населения будет снижать благосостояние, однако только к середине XIX века количество статистических данных, собранных в Европе и США, стало достаточным для создания *науки о населении*. Прародитель демографии **Жан-Поль Ашиль Гийяр** (Guillard) (1799–1896) впервые использовал слово «демография» для обозначения новой науки в 1855 году.

Соревнование между Англией и Францией, обостренное Французской революцией и погружением Европы в войну после 1793 года, заставило английское общество задуматься о количестве людей, годных к военной службе, в последнем десятилетии XVIII века.

На протяжении Наполеоновских войн философ-утилитарист **Иеремия Бентам** (1748–1832) обнаружил, что правительство не знает, сколько бедняков получали пособие и даже не знает о количестве денег, находящихся в обороте.



Лондонское статистическое общество

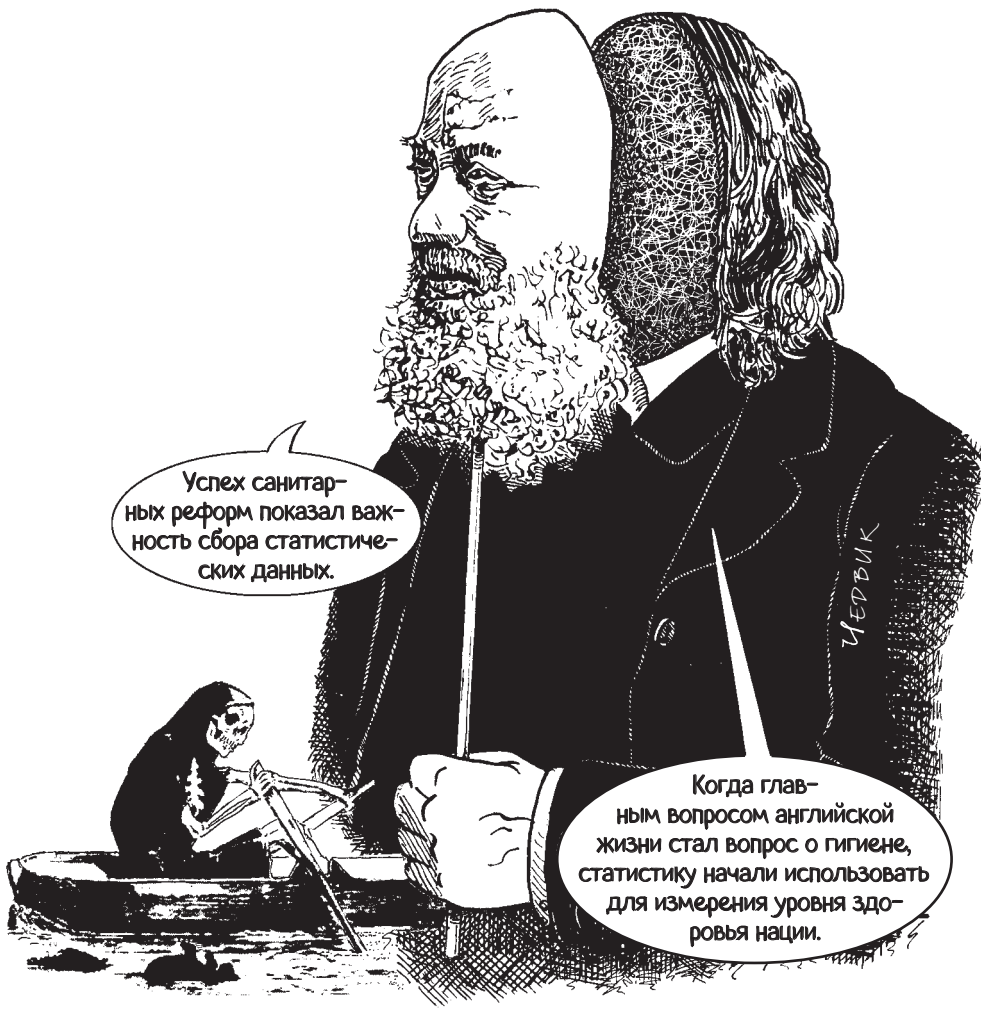
Отсутствие официальных учреждений послужило толчком к созданию Лондонского статистического общества (в настоящее время известного как Королевское статистическое общество) в 1834 году. Мальтус совместно с бельгийским статистиком и метеорологом **Адольфом Кетле** (1796–1874) и **Чарльзом Бэббиджем** (1791–1871), который разработал первую вычислительную машину (предшественник компьютера), приложили усилия к созданию такого общества.

Управление записи актов гражданского состояния (УЗАГС) было создано, предоставляя Англии и Уэльсу систему по сбору демографических данных, уникальную для Европы того времени. Первая полноценная перепись была проведена в Англии в 1851 году и включала информацию о возрасте, поле, профессии и месте рождения, а также данные о глухоте и слепоте.

Эдвин Чедвик и санитарные реформы

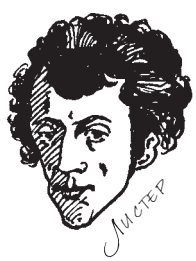
Первая перепись предоставила детальную информацию о количестве смертей от болезней и помогла осознанию ужасающих санитарных условий в городах. Перенаселение часто приводило к жилищным условиям, в которых не предусматривалось адекватной вентиляции и гигиены. Сточные колодцы были переполнены, а канализации вели напрямую в реки, увеличивая опасности для здоровья окружающих.

Главной фигурой санитарной реформы и использования для нее статистических данных был либерально настроенный **Эдвин Чедвик** (1800–1890), который участвовал в реорганизации государственной помощи беднякам и нуждающимся.



Уильям Фарр и демографическая статистика

После того как УЗАКС было создано, Чедвик рекомендовал назначить начальника службы регистрации актов гражданского состояния для фиксирования рождений и смертей. Парламент создал такую должность, и **Томас Генри Листер** (Lister) (1800–1842), родственник одного широко уважаемого британского министра, был на нее назначен.



Однако было необходимо создать команду, которая бы занималась сбором и обработкой статистических данных, и Листер пригласил на должность **Уильяма Фарра** (1807–1893) для анализа статистики, так как он был единственным врачом, который уделял внимание демографической статистике.

Работа Фарра в качестве суперинтенданта по статистике при начальнике службы регистрации в 1839 году была поворотной точкой в развитии английской профилактической медицины и медицинской статистики. Его способы работы с демографической статистикой предоставили модель для всех остальных стран. Вместе с **Томасом Роу Эдмондсом** (Edmunds) (1803–1899) они создали современную науку, называемую демографической

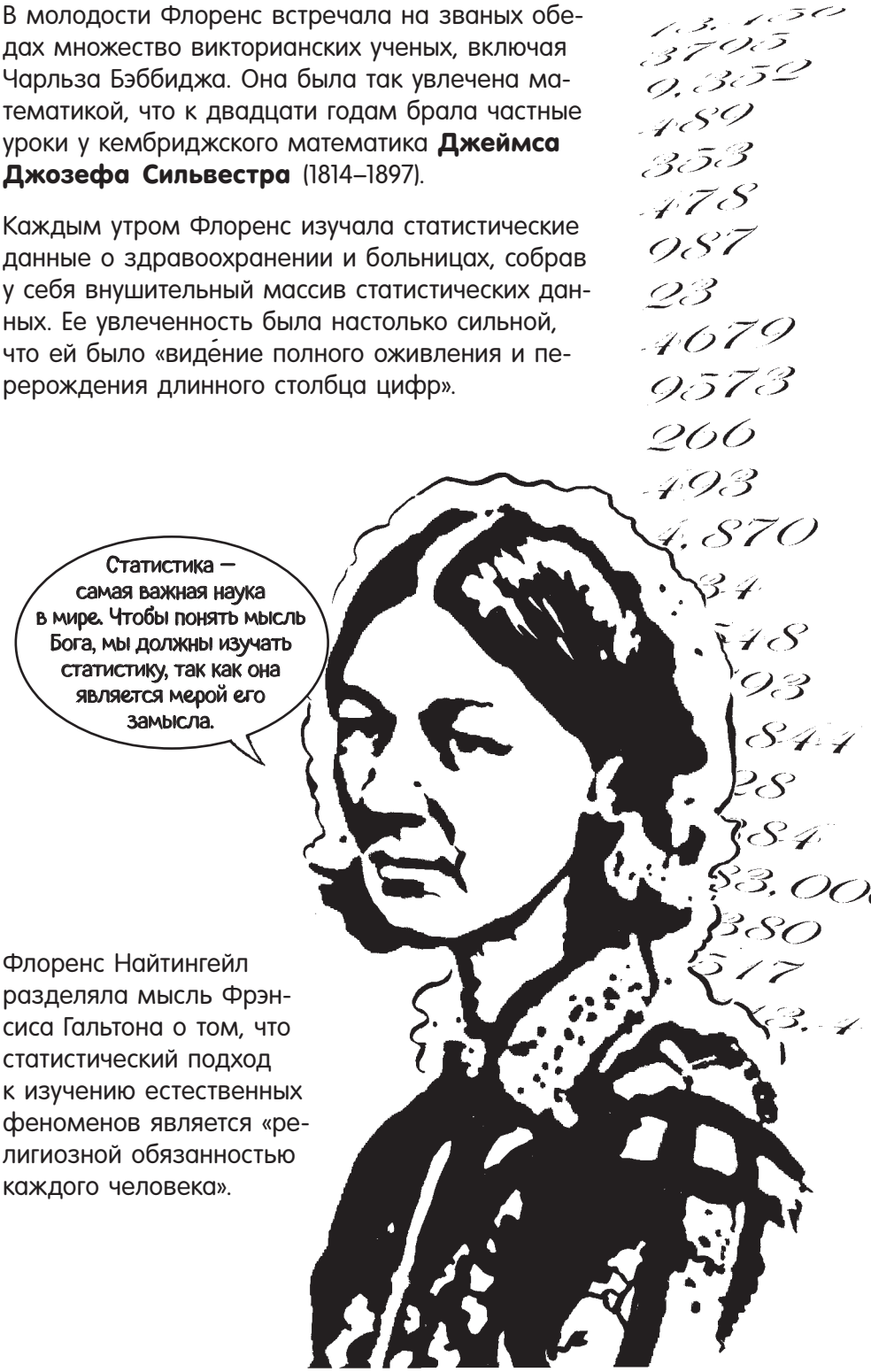
Флоренс Найтингейл: увлеченный статистик

Статистические труды Фарра и Кетле вдохновили Флоренс Найтингейл (Nightingale) (1820–1910), одну из наиболее ярких представителей викторианцев, известную многим как «леди со светильником», которая сделала профессию сестры милосердия уважаемой. Тем не менее нам мало известно о ее роли «увлеченного ученого-статистика», — эпитет, данный в 1913 году ее первым биографом, сэром Эдвардом Куком.

Используя методы и идеи статистиков-викторианцев среднего периода (mid-Victorian), Найтингейл убедила многих государственных чиновников в важности того опыта, который она получила во время Крымской войны, и показала, что смертность в войсках может быть снижена.

В молодости Флоренс встречала на званых обедах множество викторианских ученых, включая Чарльза Бэббиджа. Она была так увлечена математикой, что к двадцати годам брала частные уроки у кембриджского математика **Джеймса Джозефа Сильвестра** (1814–1897).

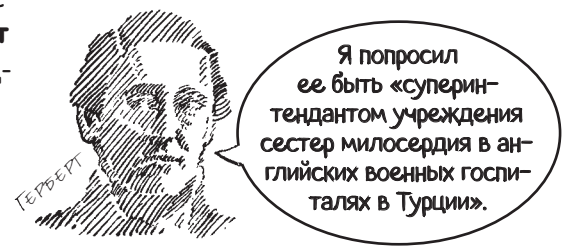
Каждым утром Флоренс изучала статистические данные о здравоохранении и больницах, собрав у себя внушительный массив статистических данных. Ее увлеченность была настолько сильной, что ей было «виде́ние полного оживления и перерождения длинного столбца цифр».



Флоренс Найтингейл разделяла мысль Фрэнсиса Гальтона о том, что статистический подход к изучению естественных феноменов является «религиозной обязанностью каждого человека».

Статистика Крымской войны

В 1854 году близкий друг Флоренс военный министр **Сидни Герберт** (1810–1861) обратился к ней с предложением.



Вместе с группой из 38 сестер милосердия она должна была заботиться о британских военных, сражавшихся на Крымской войне.

Ее связи в правительстве и годы, потраченные на защиту профессии сестры милосердия, позволили состояться этому исключительному назначению. До этого времени женщинам не позволялось состоять на службе.

В газете *«Таймс»* Герберт отвечал на гнев людей, вызванный отчетами о войне…

Герберт надеялся, что присутствие Найтингейл успокоит общественность. Читатели «Таймса» пожертвовали ей 7000 фунтов на личные нужды, которые в конечном счете пошли на улучшение полевых госпиталей, в то же время вызвав зависть со стороны других военных врачей и офицеров.

Как только Найтингейл прибыла в Крым, она обнаружила абсолютный хаос, царивший в госпитале в Скутари: не было мебели, еды, посуды, простыней и кроватей, зато повсюду были крысы и блохи. В госпитале она раздобыла чашки для чая, которые использовались солдатами для умывания, еды и питья.

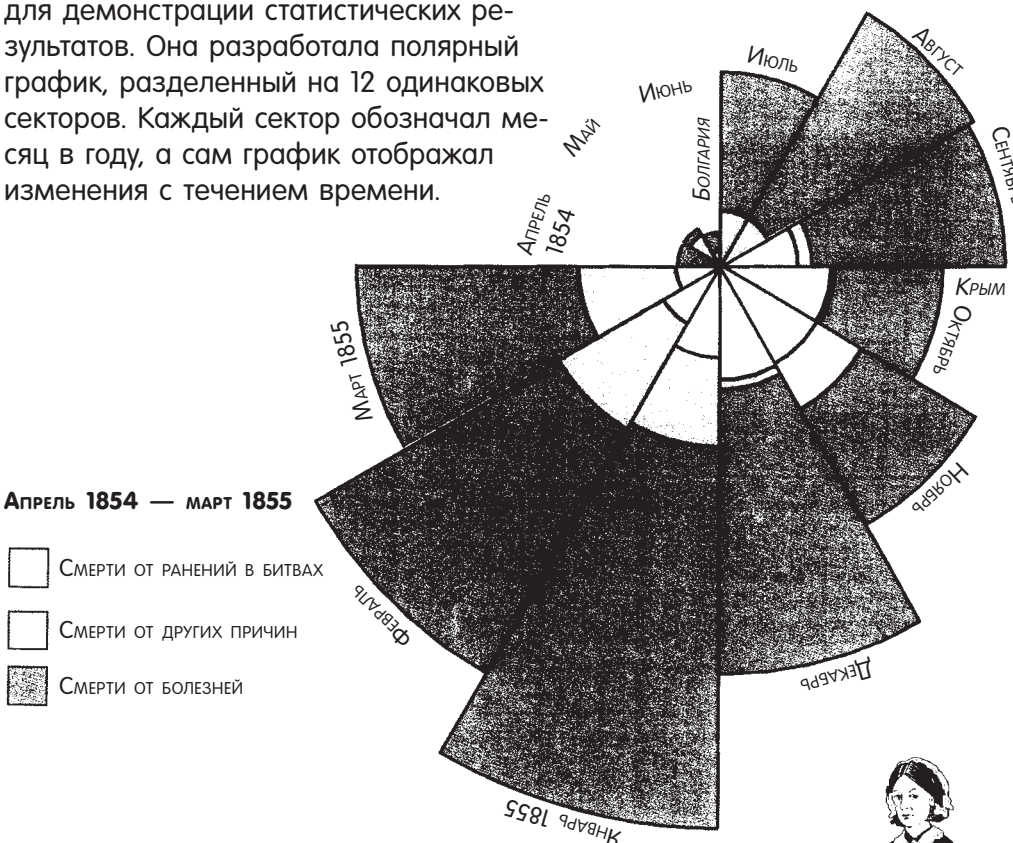
Она была единственным человеком с финансами и авторитетом, который мог бы исправить эту чудовищную ситуацию. Найтингейл заказала посуду, рубашки, простыни, покрывала, сумки для матрацев, столы для операций, ширмы и льняную ткань для перевязок. Вскоре она организовала прачечную и кухню, а бóльшую часть еды поставляла компания Fortnum & Mason.

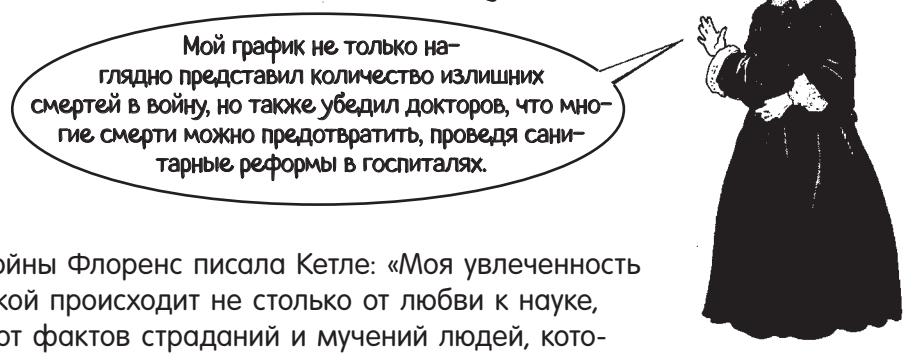
Статистика смертности в Крыму

Найтингейл была очень возмущена небрежностью, с которой велась статистика в военных госпиталях. Связь между госпиталями была ужасной, и не было единой формы отчетности. Каждый госпиталь использовал свою классификацию болезней, ведя отчетность в различных формах, что делало невозможным сопоставления. Даже количество смертей подсчитывалось невнимательно: сотни мужчин были похоронены, однако их смерти не были зафиксированы.

Полярный график

Несмотря на то, что различные статистики-демографы XIX века использовали целый ряд графиков и таблиц для своих результатов, Найтингейл способствовала популяризации наглядных диаграмм для демонстрации статистических результатов. Она разработала полярный график, разделенный на 12 одинаковых секторов. Каждый сектор обозначал месяц в году, а сам график отображал изменения с течением времени.





После войны Флоренс писала Кетле: «Моя увлеченность статистикой происходит не столько от любви к науке, сколько от фактов страданий и мучений людей, которые я так часто наблюдала из-за безразличия законов и правительств».

Вероятность

Как статистики XIX века сводили данные к чему-то более управляемому и, значит, полезному? Хотя данные уже записывались в диаграммы и таблицы, до конца XIX века двумя основными статистическими инструментами оставались вероятность и средние значения.

Вероятность является одним из старейших статистических понятий, его использовали еще в начале XIV века как метод при решении задач, основанных на случае.

Есть несколько подходов к понятию вероятности:

1. **Субъективный.**
2. **Игры со случайным исходом.**
3. **Математический.**
4. **На основе относительной частоты событий (частости).**
5. **Байесовский.**

Вместе с шестью основными распределениями вероятностей:

1. **Биномиальное распределение.**
2. **Распределение Пуассона.**
3. **Нормальное распределение.**
4. **Распределение хи-квадрат.**
5. **t-распределение.**
6. **F-распределение.**

Первые три распределения мы рассмотрим на с. 46–50. Последние три распределения используются для определения статистической значимости. Значимость по тесту хи-квадрат рассмотрена на с. 153–156, t- и F-распределения будут рассмотрены на с. 165 и 170 соответственно.

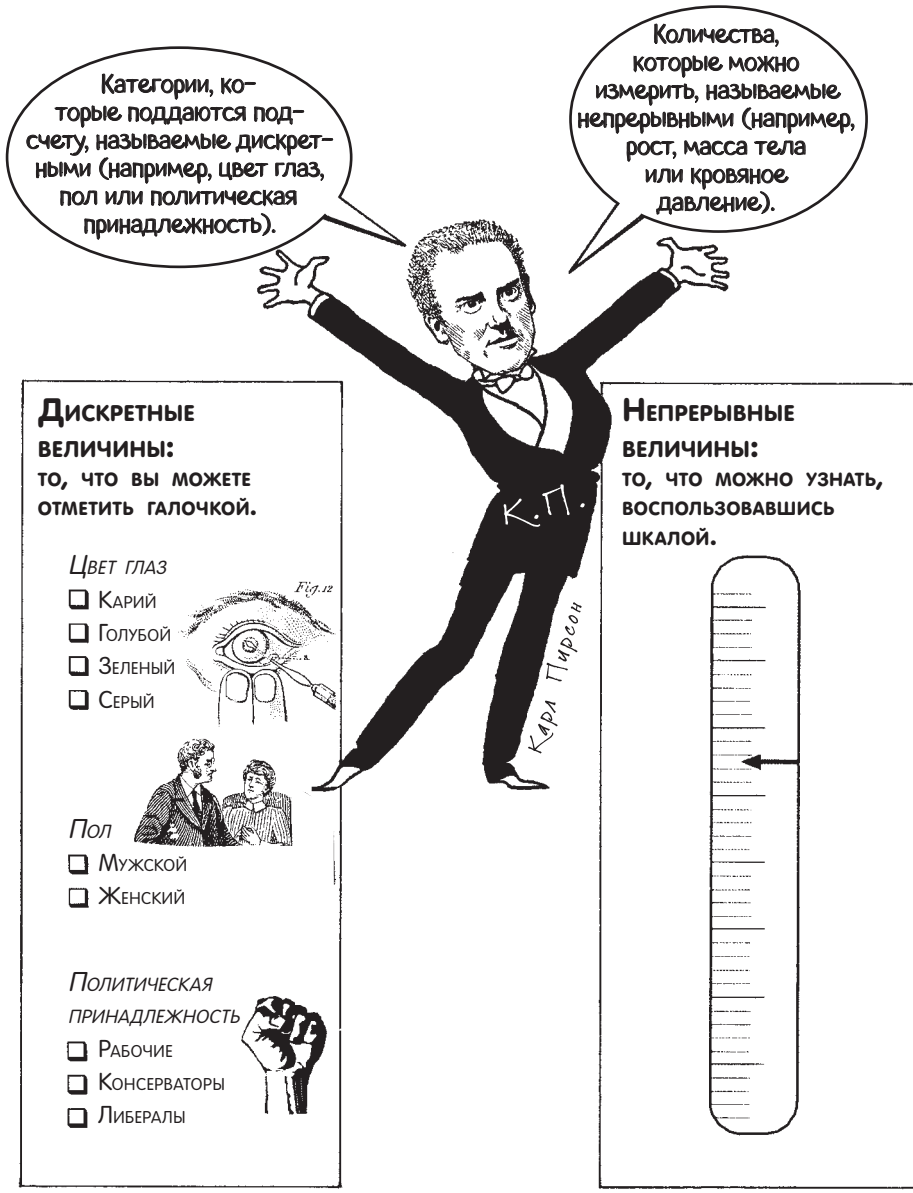
Есть два типа статистического распределения: **распределение вероятностей**, которое описывает возможности того или иного исхода в выборке, и частоту, с которой каждый из исходов будет иметь место; и **частотные распределения** (см. с. 74, 76, 79–85), которые описывают частоту возникновения каждого исхода.

Статистики используют распределения вероятностей для интерпретации набора данных, которые анализируются различными статистическими методами. Частотные распределения помогают перевести большое количество чисел и групп чисел в более удобную для работы форму и показывают, как часто встречается тот или иной исход каждого события в группе.

Переменные величины

Переменные величины являются характеристикой индивида или системы, которые можно измерить и подсчитать. Они могут изменяться во времени или от индивида к индивиду.

Переменные величины можно разделить на два типа:



Эти переменные величины можно классифицировать более детально, о чем мы поговорим позднее.

Субъективный подход к вероятности заключается в вере (belief) в наиболее рациональный исход.

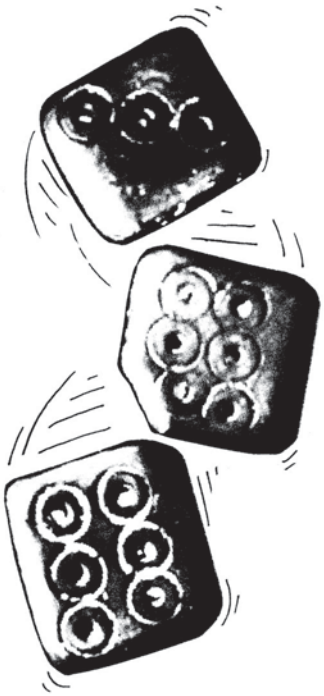
Возможные исходы часто отражают личное мнение. Два человека могут высказать разные предположения о вероятности исхода (наступления события), но нет никакой объективной процедуры, которая докажет, что один прав, а другой нет.

Азартная игра, или **пари**, определяется как схема ставок, основанная на том, какую вероятность предсказывает участник игры. Идея состоит в том, чтобы обнаружить вероятность, которую определяет сам участник, а не вероятность, проистекающую из внешнего мира. Проблема заключается в том, что люди с одинаковыми знаниями и навыками приходят к разным ответам.

Игры со случайным исходом

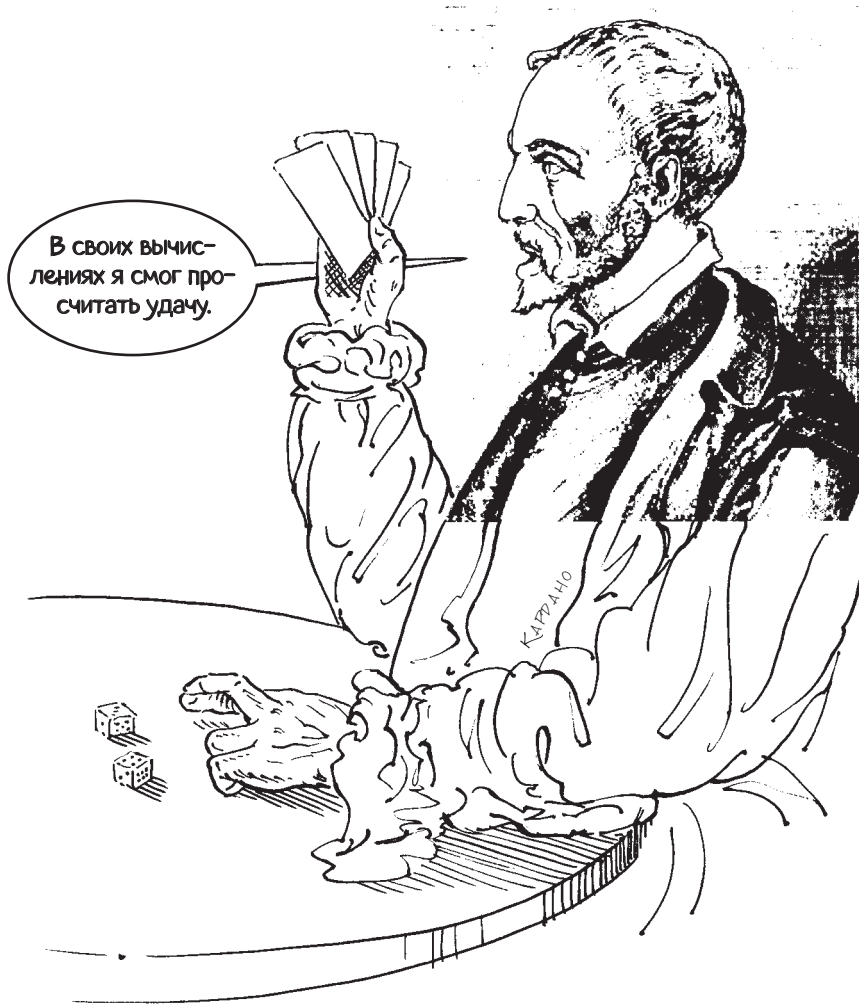
Игры, основанные на случайности, появились в тот момент, когда человек смог бросить кости. Согласно археологическим исследованиям Северного Ирака, человек играл в подобные игры в Месопотамии еще до начала III тысячелетия до н. э. Кости также использовались во времена 18-й династии в Египте (1400 год до н. э.).

Первые кости были сделаны из длинных костей животных, выточенных в форме квадрата. Астрагал (бедренная кость небольшого размера) обычно использовался в качестве игровых костей, которые бросали древние греки, а затем римляне.



Данте Алигьери
(1265–1321)

Итальянский физик и математик эпохи Возрождения **Джероламо Кардано** (1501–1576) был заядлым игроком, который часто зарабатывал игрой себе на жизнь. Он написал одну из первых работ по теории вероятности: «Liber de Ludo Aleae» («Об азартных играх»), опубликованную посмертно в 1633 году. Эта книга служила руководством для игроков.



Однако удача исчезла в XVII веке, когда возникла классическая теория вероятностей. Согласно этой теории, весь набор возможных исходов следует включать в математическую вероятность. Поэтому даже удачливому Кардано пришлось соответствовать этим математическим требованиям.

Де Муавр и азартные игры в Сохо*

В 1718 году французский математик **Абрахам де Муавр** (1667–1754) написал труд по теории вероятностей *Doctrine of Chance: or A Method of Calculating the Probabilities of Events in Play* («Учение о случае, или Метод вычисления вероятностей событий в играх»). Этот труд основывался на проблемах преимущества игроков и размерах их ставок в играх. Как и труд Кардано, работа де Муавра служила руководством для игроков.

Он был вынужден перебраться из Франции в Англию в 1685 году из-за отмены Людовиком XIV Нантского эдикта, который заставил сотни тысяч французских протестантов бежать из Франции.

Во время пребывания в Лондоне де Муавр познакомился с Эдмондом Галлеем и Исааком Ньютоном, а также был избран членом Лондонского королевского общества в возрасте тридцати лет.

* Сохо — район французских и итальянских ресторанчиков и ночных клубов в Лондоне, район богемы. — *Прим. науч. ред.*

Математическая теория вероятностей

К концу XVII века идеи вероятности в комбинаторике (раздел математики, изучающий перестановки и комбинации) были применены к играм со случайным исходом такими учеными, как:

…но они не знали, как подсчитать неопределенность.

Математическая теория вероятностей дала статистикам инструмент, который избавлял от сложностей, показывал, какие закономерности можно вывести из случая, и даже сводил сам случай к набору законов.

Этот подход описывал долгосрочную закономерность в случайных событиях, а также вычислял отношение числа благоприятных случаев к числу возможных случаев:

$$\frac{\textit{благоприятные случаи}}{\textit{возможные случаи}}$$

В таком теоретическом подходе не учитывались реальные объекты, нужно было лишь предположить ряд гипотетических условий, а затем подсчитать вероятность, используя биномиальное распределение (см. с. 46–48).

Такое развитие математики, возникшее в XVII веке, послужило основанием для формальной теории в начале XVIII века, однако использование вероятности в статистике началось только в конце XIX века.

Частость события

Частость события — это подход, который позволяет производить формальные выражения вероятности (P, A) о недостоверных событиях, где P — это вероятность недостоверного события A. Следовательно, вероятность возникновения события пропорциональна числу возникновения таких же событий на длинном временно́м интервале.

Это более научный и объективный подход, нежели другие типы вероятности, используемый при исследовании внешнего мира и реально существующих объектов. Можно подбросить монету 100 раз, записать количество выпавших орлов и решек и получить требуемое соотношение, разделив количество выпавших орлов на общее количество бросков.

В своих ранних лекциях по статистике Карл Пирсон разбрасывал по полу аудитории сотни монет и просил студентов собрать их и отсортировать согласно выпавшим сторонам монеты.

Но как мы можем понять, сколько раз стоит подбросить монету (или бросить кости) для того, чтобы провести качественный эксперимент? Если вы будете подбрасывать монету и получите 60 орлов и 40 решек, у вас вряд ли получится повторить результат. Вероятность будет всегда изменяться, и к тому моменту, как она станет устойчиво постоянной (stable), ваша монета будет затерта с обеих сторон.

Выходом из такого затруднения является коэффициент относительной частоты событий, или коэффициент частости. Он получается как отношение числа испытаний, в которых данное событие появилось, к общему числу фактически проведенных испытаний.

Байесовский подход

Математик **Томас Байес** (или **Бейес** (Bayes)) (1702–1761) впервые использовал вероятность индуктивным путем, учреждая математическую основу для выведения вероятности. Однако термин «Байесовский» применительно к статистике вошел в обиход только в 1950 году.

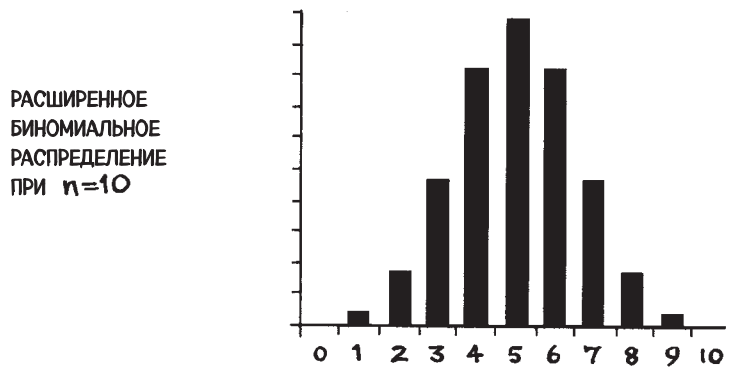
Теорема Байеса — это формула, которая показывает, как существующие разумные предположения (beliefs), формально записанные как распределения вероятности, изменяются под влиянием новой информации.

Пример.

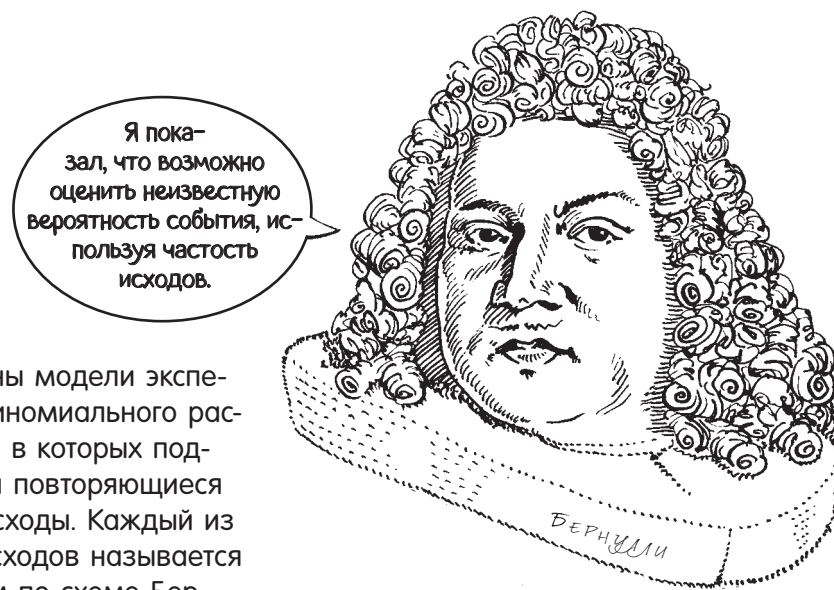
Теорема Байеса может быть использована в диагностике, которую проводят терапевты и лечащие врачи. Эти доктора обычно начинают с априорного предположения о том, болен пациент или нет конкретной болезнью (основываясь на знании симптомов пациента и распространенности заболевания), и это знание может быть изменено или улучшено посредством результатов анализов пациента.

Распределение вероятностей

Биномиальное распределение — это дискретное распределение вероятностей, показывающее вероятность двух исходов события, которое может произойти, а может и не произойти. Оно описывает возможное количество случаев, в которых случается событие на протяжении последовательности наблюдений. Например, оно дает возможность вычислить вероятность выпадения пяти решек при подбрасывании десяти монет.



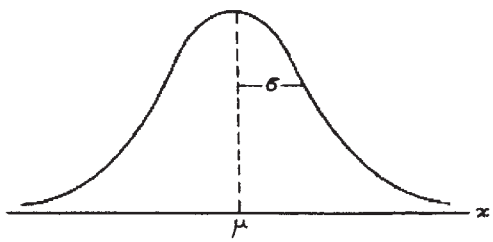
Это распределение было введено в науку швейцарским математиком **Якобом Бернулли** (1655–1705), прославленная работа которого, — *Ars conjectandi* («Искусство предположения»), — была опубликована посмертно в 1713 году. Эта работа знаменует начало математической теории вероятностей.



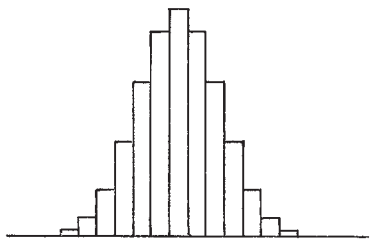
Были созданы модели экспериментов биномиального распределения, в которых подсчитываются повторяющиеся двоичные исходы. Каждый из двоичных исходов называется «испытанием по схеме Бернулли».

Биномиальное распределение $(p+q)^n$ определяется количеством наблюдений **n** и вероятностью появления события, обозначаемую **p+q** (два возможных исхода).

Оно предоставляет модель для измерения различных вероятностей исходов, которые могут произойти. Для определения вероятности каждого исхода биномиальное распределение необходимо расширить количеством наблюдений, что достигается путем возведения **p+q** в степень **n**.



НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ



БИНОМИАЛЬНОЕ РАСПРЕДЕЛЕНИЕ, ПОСТОЯННО СТРЕМЯЩЕЕСЯ К НОРМАЛЬНОМУ РАСПРЕДЕЛЕНИЮ.

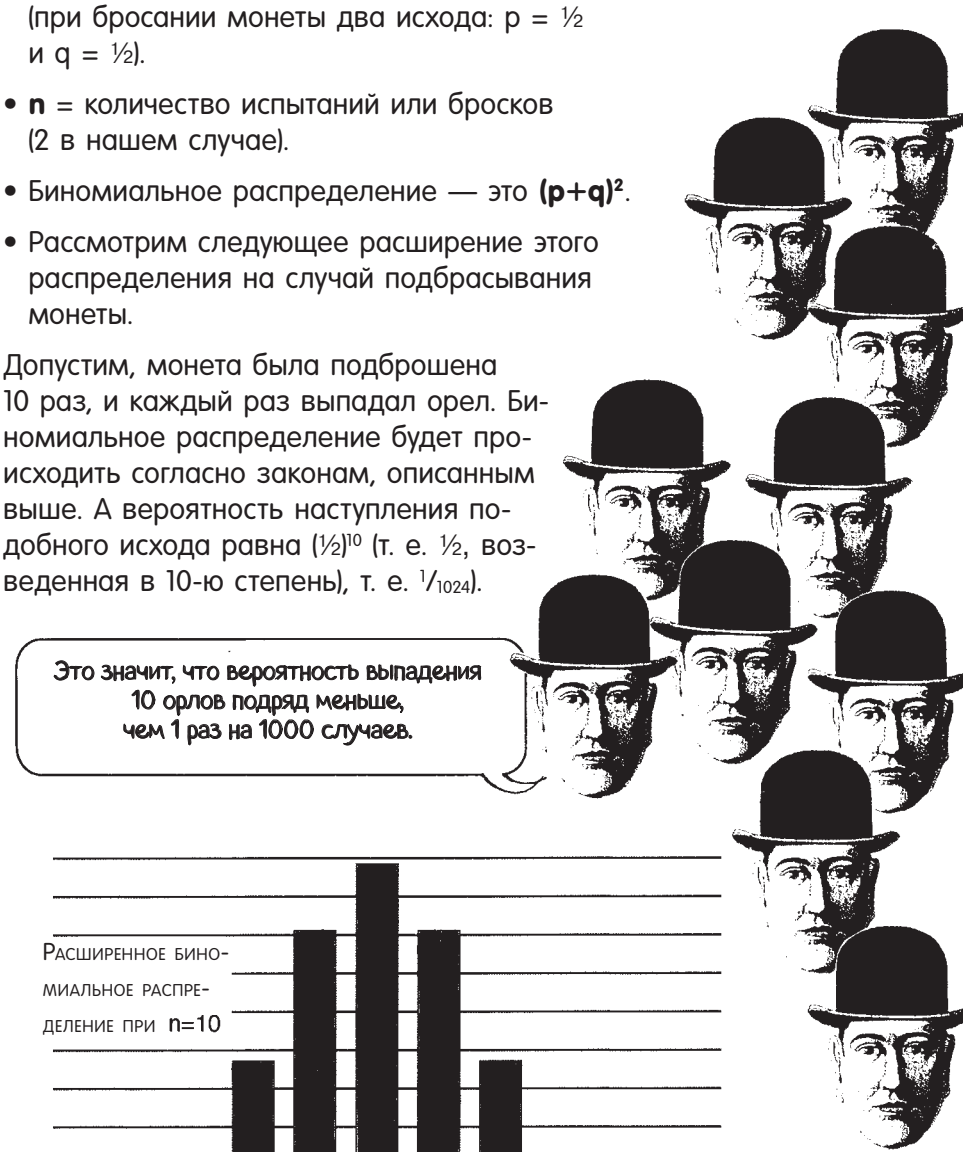
Подобные вероятностные распределения соотносятся с различными типами переменных величин. Дискретные вероятностные распределения, такие как биномиальное, используют дискретные данные (например, орел или решка в подбрасывании монеты), в то время как непрерывные распределения, например нормальное (или Гауссово), используют непрерывные величины, такие как рост и масса тела.

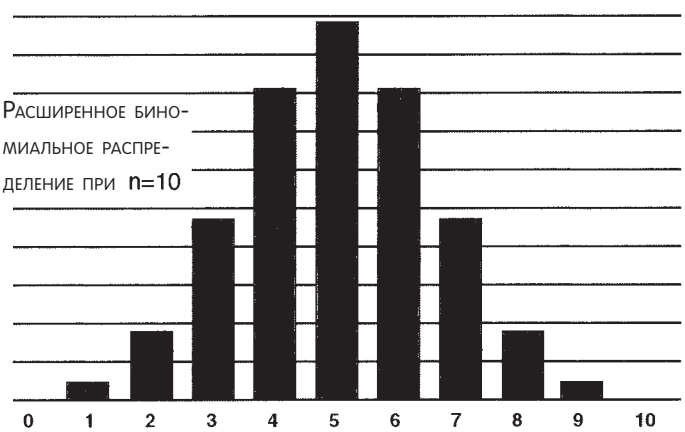
В следующем примере бросания монеты количество наблюдений **n=2**, количество исходов 2 (орел или решка). Для проверки идеальной монеты биномиальное распределение следует расширить для приведения в соответствие количества подбрасываний монеты.

Расширим биномиальное распределение $\mathbf{(p+q)^n}$, возводя p+q в степень n (что означает умножение числа само на себя).

- **p** и **q** в сумме должны давать 1 (при бросании монеты два исхода: p = ½ и q = ½).
- **n** = количество испытаний или бросков (2 в нашем случае).
- Биномиальное распределение — это $\mathbf{(p+q)^2}$.
- Рассмотрим следующее расширение этого распределения на случай подбрасывания монеты.

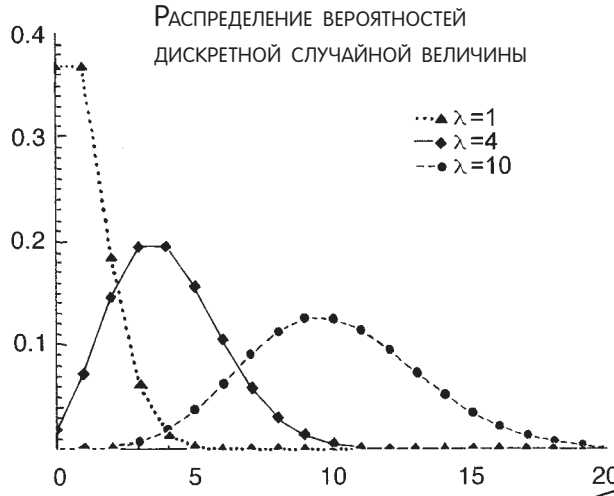
Допустим, монета была подброшена 10 раз, и каждый раз выпадал орел. Биномиальное распределение будет происходить согласно законам, описанным выше. А вероятность наступления подобного исхода равна $(½)^{10}$ (т. е. ½, возведенная в 10-ю степень), т. е. $^1/_{1024}$).





Распределение Пуассона

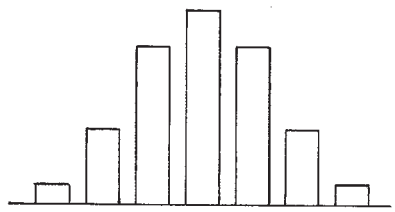
Распределение Пуассона, открытое **Симеоном Дени Пуассоном** (1781–1840) — это дискретное распределение вероятностей, используемое для описания появления неблагоприятных исходов при большом количестве независимых и повторяющихся испытаний. Такое распределение является хорошим приближением биномиального распределения, если вероятность низкая, а число испытаний велико.



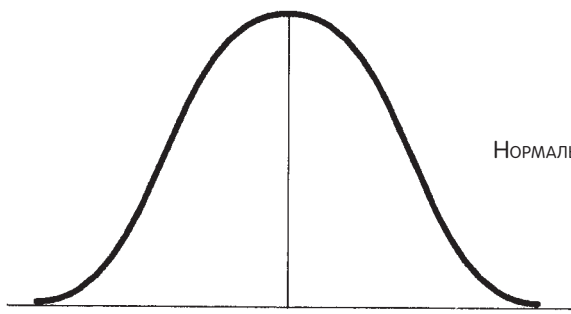
При анализе статистики смертности зачастую используется распределение Пуассона в предположении, что смерти среди населения от большинства заболеваний возникают независимо друг от друга и обладают свойством случайной величины.

Нормальное распределение

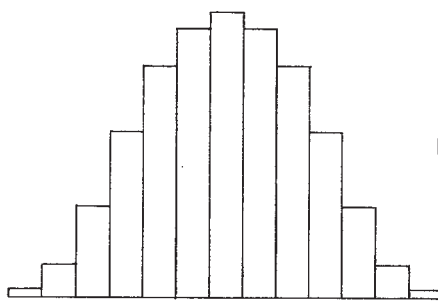
Нормальное распределение — это непрерывное распределение, и оно связано с биномиальным. По мере того как **n** стремится к бесконечности, биномиальное в предельном случае достигает нормального распределения. На графике это будет выглядеть как бесконечное количество бесконечно малых прямоугольников, и тогда биномиальное распределение станет нормальным.



Биномиальное распределение



Нормальное распределение



Биномиальное, которое стремится к нормальному

График также известен как график кривой нормального распределения, иногда (не вполне точно) называемого Гауссовым распределением. Такое распределение долго использовалось как мерило и критерий для сравнения с другими типами статистических распределений. Оно играет ключевую роль в современной статистике, так как позволяет статистикам интерпретировать данные, используя различные статистические методы, которые зачастую моделируются на основе нормального распределения.

Астрономические наблюдения

Идея кривой нормального распределения берет свое начало в вычислениях комбинаций наблюдений астрономов. Они использовали «закон ошибок» (т. е. кривую нормального распределения) для объединения линейных уравнений своих наблюдений в астрономии и геодезии*.

Астрономические методы зачастую были процедурами с узкой областью применения. Они имели косвенное отношение к формальным моделям вероятности, требовали взаимодействия группы ученых. Но когда математические статистики стали разрабатывать статистические методы, это позволило анализировать астрономические данные в одиночку.

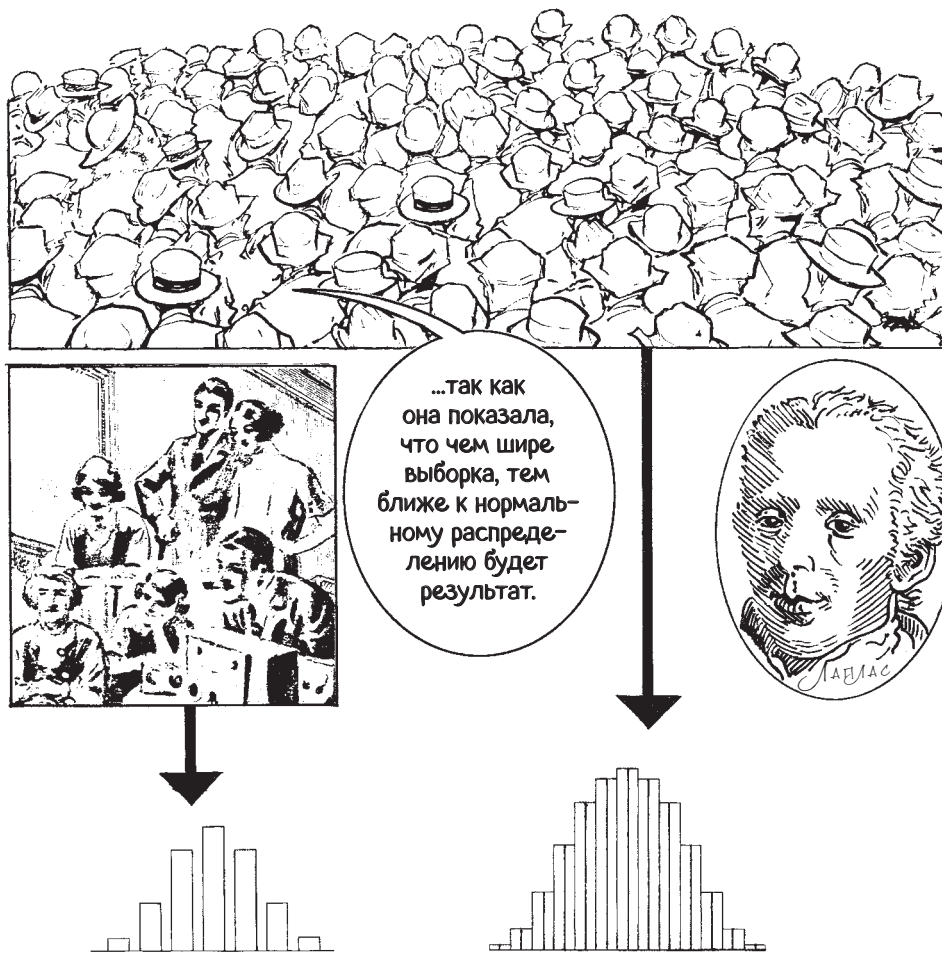
Труд де Муавра об играх со случайным исходом и его использование биномиальной теоремы предоставили в 1733 году первую известную кривую нормального распределения, которую сперва называли «законом ошибок». Он также составил первую таблицу вероятностей для нормального распределения.

* Наука о форме и областях Земли.

Центральная предельная теорема

Французский математик и астроном **Пьер-Симон Лаплас** (1749–1827) занимался усовершенствованием теории вероятностей в качестве инструмента, который может снизить и измерить недостоверность данных. К 1789 году он понял, что на измерения оказывают влияние множество независимых друг от друга небольших ошибок, и показал, что закон ошибок может быть выведен математически. Затем он внес свой главный вклад в статистику, написав в 1810 году труд о **Центральной предельной теореме**.

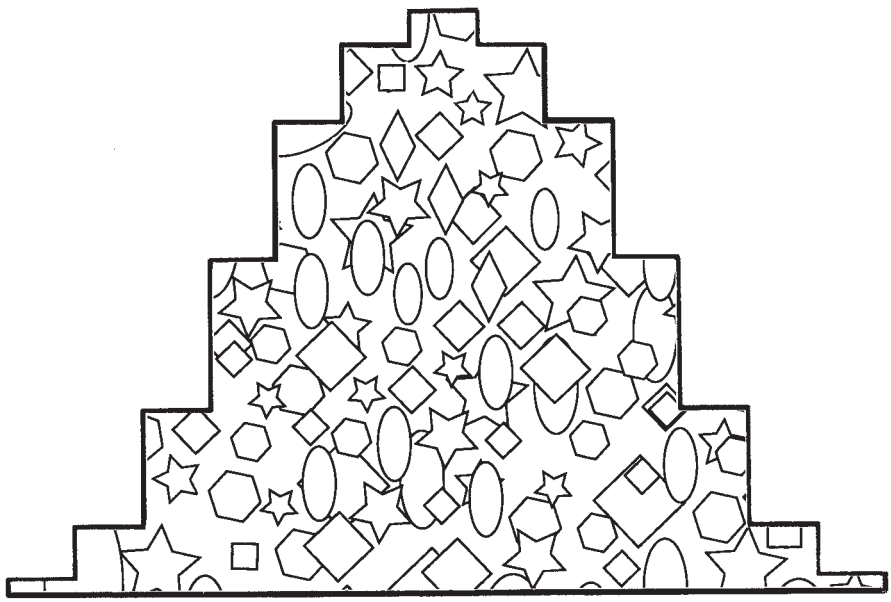
Эта теорема была одним из крупнейших достижений в теории вероятностей...



Или, как сказали бы статистики: с увеличением выборки выборочное распределение средних стремится к кривой нормального распределения, вне зависимости от отклонений от нормальности в распределении населения.

Причина того, почему многие переменные величины — такие как рост или интеллект — распределены согласно нормальному распределению, кроется в Центральной предельной теореме Лапласа.

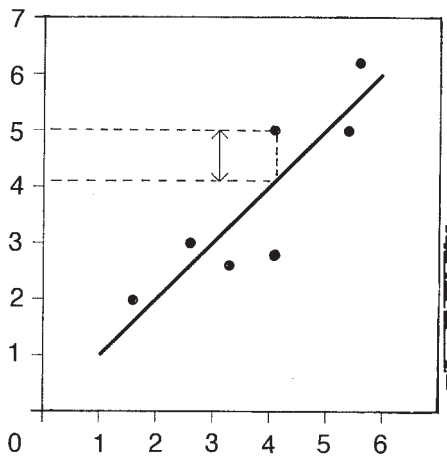
Математическое обоснование этой теоремы гласит, что данные, на которые влияют множество небольших и независимых случайных факторов (effects), будут приблизительно нормально распределены.



Гауссова кривая и принцип наименьших квадратов

Работы Лапласа оставались наиболее влиятельными в теории вероятностей до конца XIX века, хотя **Карл Фридрих Гаусс** (1777–1855) усовершенствовал идею Лапласа в ясных вероятностных формулировках. Одним из результатов данной работы было в конечном итоге (и в каком-то смысле неправильное) название «Гауссова кривая», тем не менее впервые открытая Лапласом.

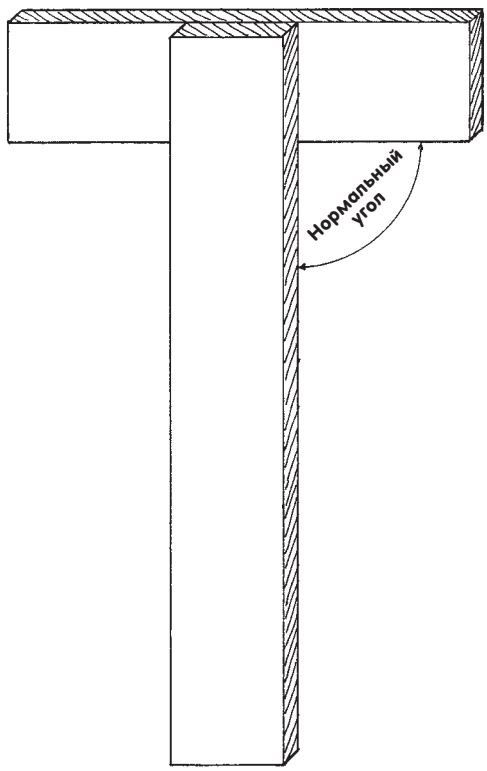
Адриен Мари Лежандр (1752–1833)

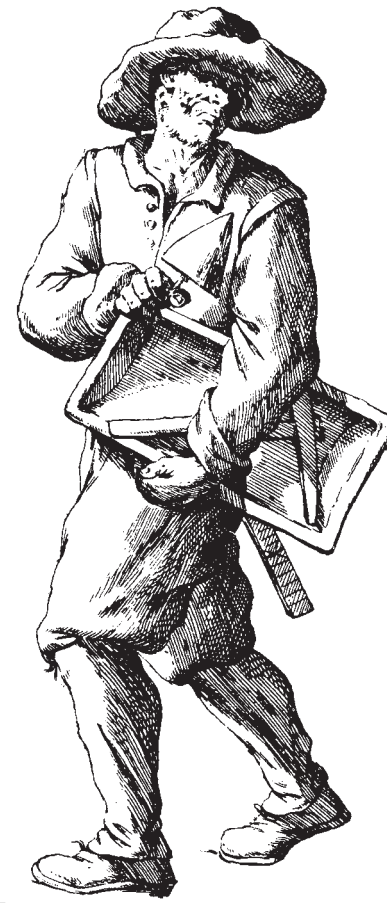


Принцип наименьших квадратов, основанный на теории ошибок, был придуман в начале XIX века такими математиками и астрономами, как Гаусс, Лаплас и Лежандр, для того, чтобы определить, например, форму Земли. Он найдет одно из своих лучших применений в статистике в конце XIX века при интерпретации статистической регрессии (см. с. 128–131).

Что такое нормальность?

Norma является латинским словом для Т-квадрата, который использовался масонами и плотниками в античности для придания своим изделиям прямоугольной формы. В результате использования ими Т-квадрата прямой угол стал известен как «нормальный угол», термин, который употреблялся в геометрии в XVII веке. Гаусс, который изучал кривую нормального распределения в 1809 году, использовал слово «норма» в алгебре в конце XVIII века.

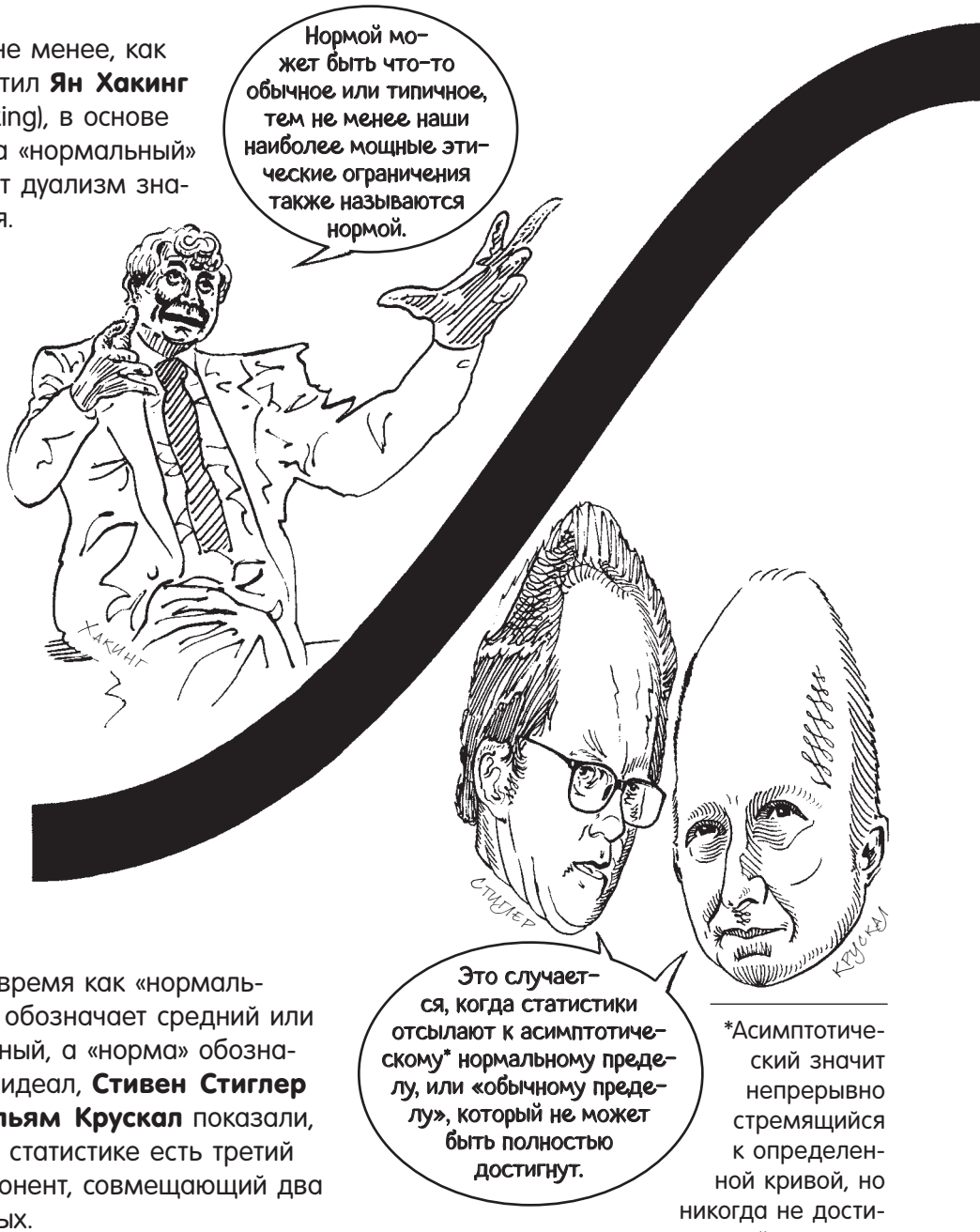




Слово «нормальный» получило распространение в XIX веке, сперва в медицинской сфере. Оно рассматривалось, как антоним слову **«патологический»**, однако вскоре стало употребляться ко всему, в особенности к людям и их поведению.

Слово «нормальный», как следствие, употреблялось для выражения того, какими в действительности *являются* вещи или какими им *следует быть*, и в результате было использовано для описания симметричного распределения в форме колокола. Такой тип распределения часто использовался астрономами с XVII века и статистиками с 1870-х годов.

Тем не менее, как заметил **Ян Хакинг** (Hacking), в основе слова «нормальный» лежит дуализм значения.

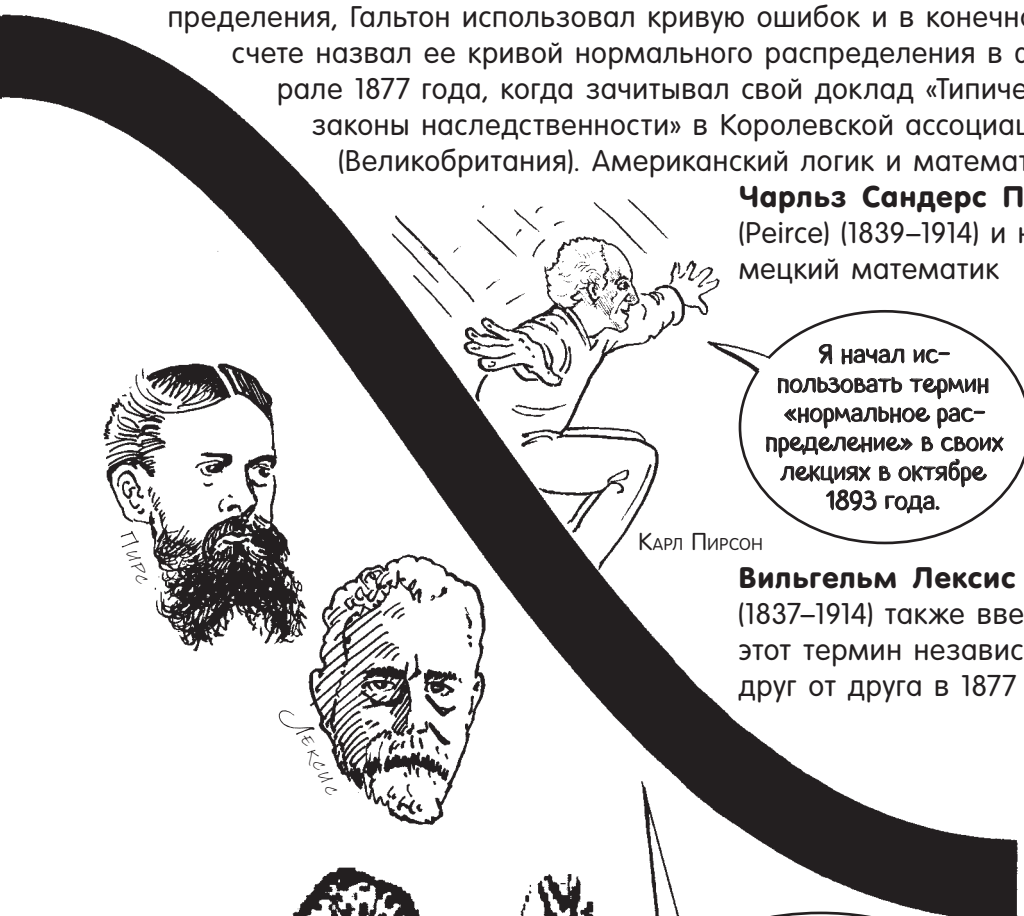


В то время как «нормальный» обозначает средний или обычный, а «норма» обозначает идеал, **Стивен Стиглер** и **Уильям Крускал** показали, что в статистике есть третий компонент, совмещающий два первых.

*Асимптотический значит непрерывно стремящийся к определенной кривой, но никогда не достигающий ее на конечных значениях переменной.

Именуемое нормальным распределением

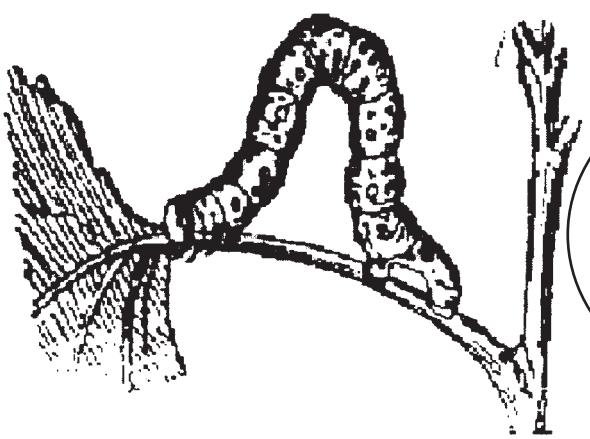
Пока Кетле использовал биномиальный закон для описания этого распределения, Гальтон использовал кривую ошибок и в конечном счете назвал ее кривой нормального распределения в феврале 1877 года, когда зачитывал свой доклад «Типические законы наследственности» в Королевской ассоциации (Великобритания). Американский логик и математик **Чарльз Сандерс Пирс** (Peirce) (1839–1914) и немецкий математик



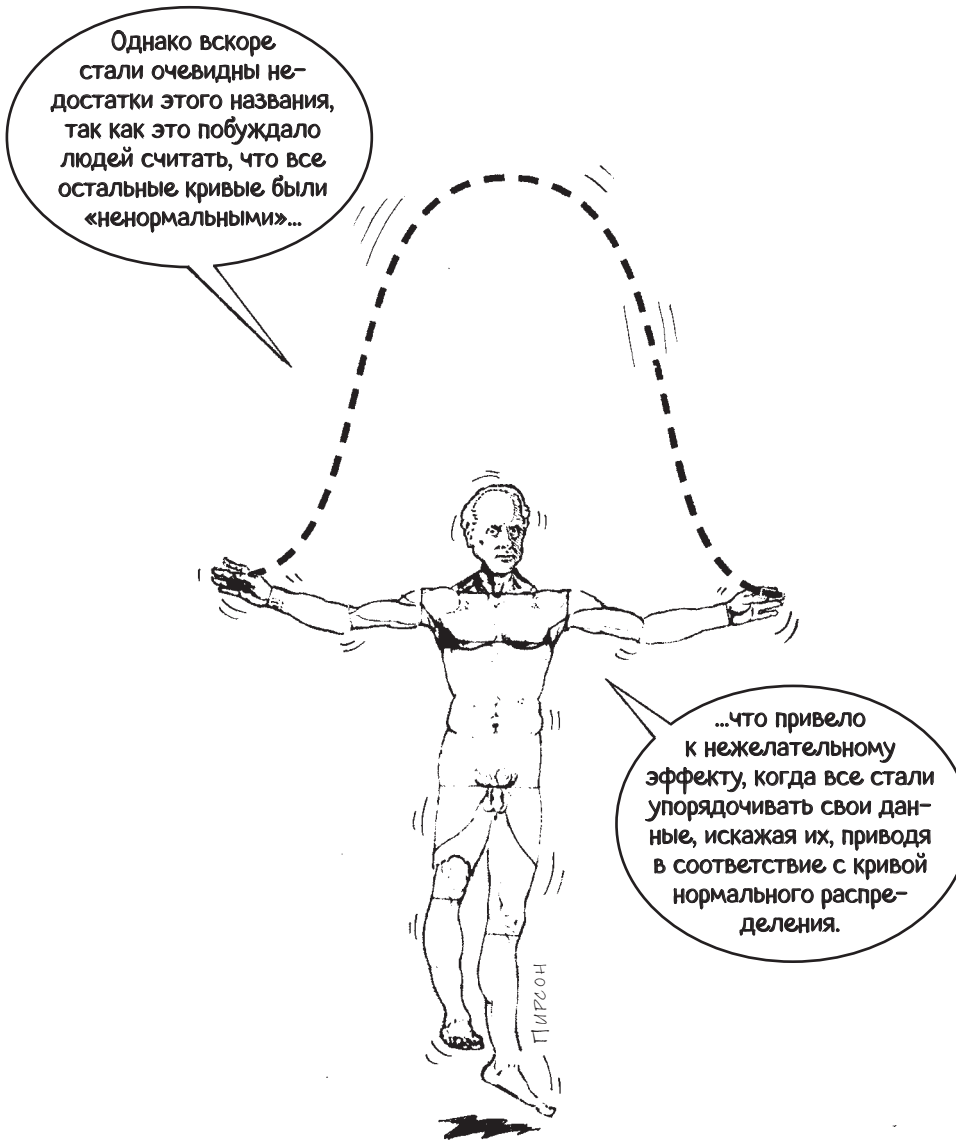
Карл Пирсон

Я начал использовать термин «нормальное распределение» в своих лекциях в октябре 1893 года.

Вильгельм Лексис (Lexis) (1837–1914) также ввели этот термин независимо друг от друга в 1877 году.



Как только я обнаружил, что Гауссова кривая была впервые открыта Лапласом, я предложил называть ее кривой Лапласа — Гаусса и в итоге начал ссылаться на нее как на кривую нормального распределения, для того, чтобы избежать международных вопросов о приоритете.



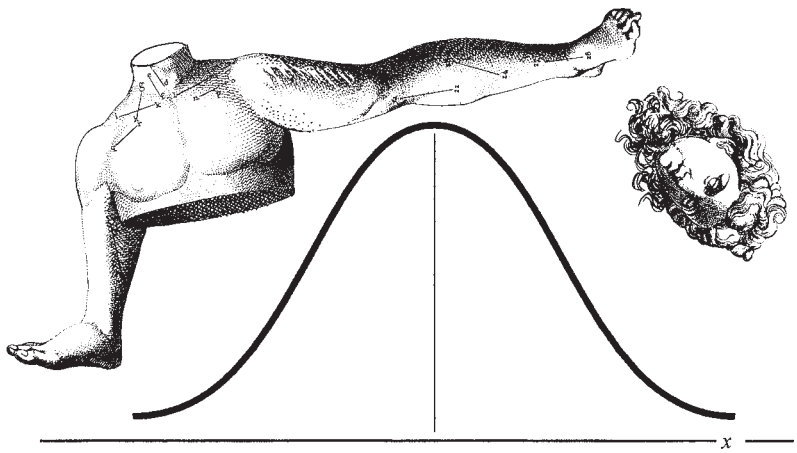
Тем не менее Пирсон был тем человеком, который популяризовал термин «нормальное распределение» в среде статистиков по всему миру.

Так что же такое нормальное распределение?

Для статистика это теоретическая конструкция, используемая для выражения того, что могло бы быть истинным в отношении собираемых данных и вероятности появления соответствующих значений с элементом случайности.

Кривая нормального распределения имеет три математических свойства:

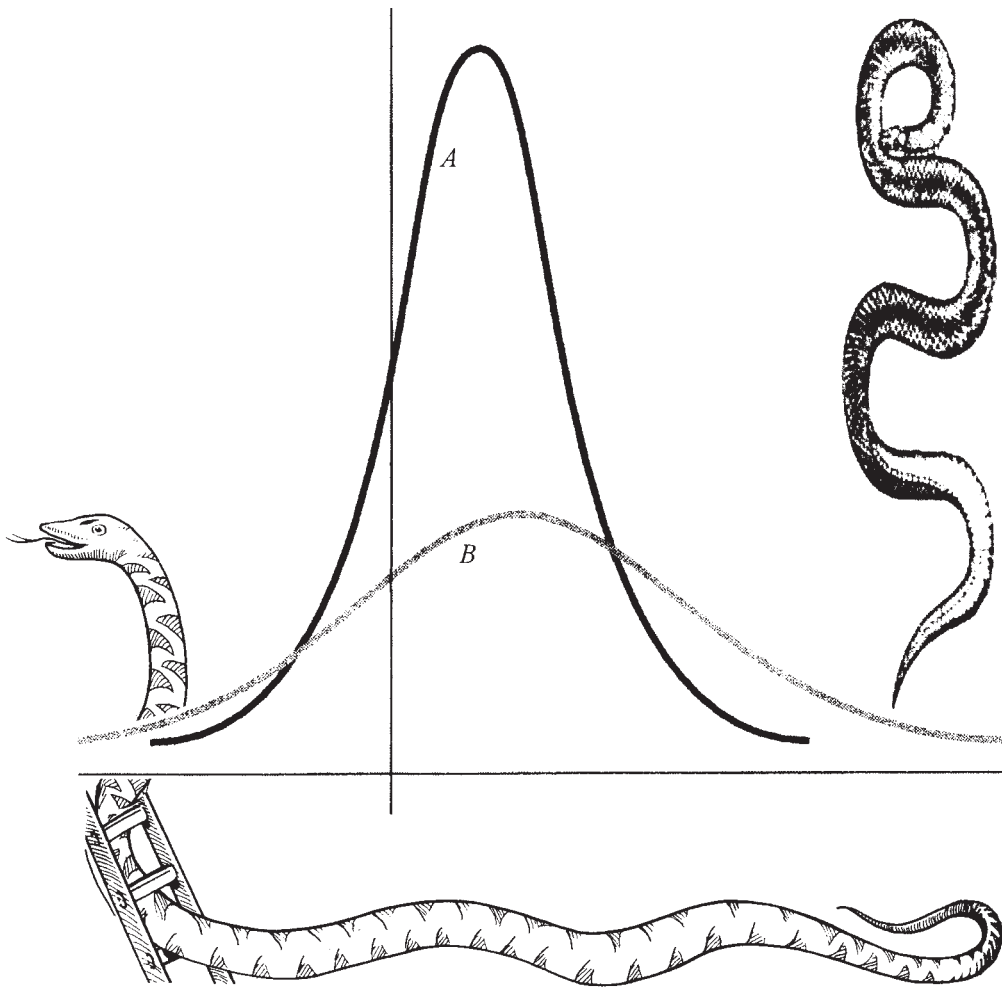
1. Она симметрична и имеет форму колокола, непрерывна и простирается от отрицательной бесконечности до положительной бесконечности.



Прямоугольное распределение также симметрично, так как имеет одинаковые частости для всех положительных значений на оси X.

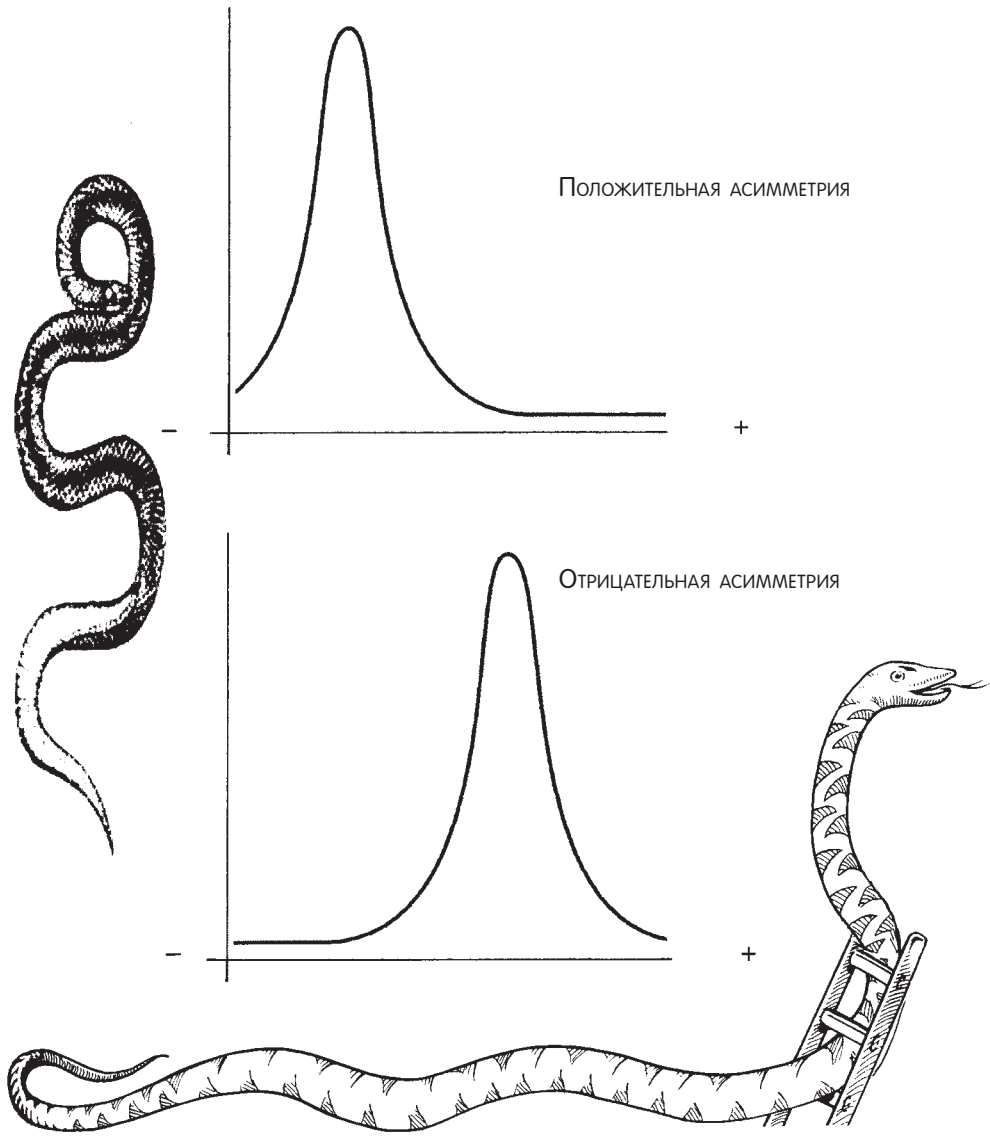
2. Среднее (см. с. 65–67) и среднеквадратическое отклонение (см. с. 99–102) определяют ее форму. Теоретическая кривая нормального распределения имеет нулевое математическое ожидание и среднеквадратическое отклонение, равное 1. Различные виды этих отклонений дают слегка различные формы кривой.

Среднее является распределением по оси X и показывает, как варьируют величины и каково рассеивание. На представленных графиках среднее значение одинаково, однако кривая B имеет бóльшую дисперсию (вариабельность), чем кривая A.



3. *Асимметрия* (скошенность) кривой нормального распределения равна 0, так как она симметрична относительно среднего значения. Если бы в распределении был перекос в левую сторону, значение асимметричности было бы отрицательным; если бы перекос был в правую сторону, значение было бы положительным.

Направление хвоста показывает, положительна или отрицательна асимметрия.



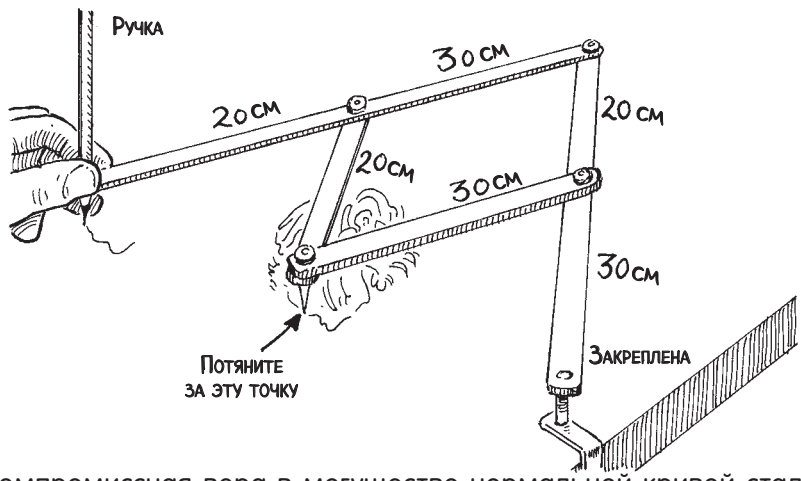
Кетлесимус

Нормальное распределение оказало сильное влияние на ряд математиков, философов и статистиков XIX века, в особенности на Адольфа Кетле (Quetelet) и Фрэнсиса Гальтона (Galton). Оба верили, что в действительности все данные должны подчиняться кривой нормального распределения.

Убежденность Кетле в том, что собранные данные могут быть сопоставлены только с кривой нормального распределения, была настолько сильна, что доктрину назвали «Кетлесимус», основываясь на том, что он преувеличивал распространенность кривой нормального распределения. Несмотря на то, что Кетле знал, что многие распределения были асимметричны, он считал, что так происходит из-за «любопытных случайных причин, действующих неравномерно в двух направлениях».

Пантограф Гальтона

Вдохновленный Кетле, Гальтон был так одержим идеей повсеместной кривой нормального распределения, что создал механическое устройство — усложненный пантограф, для того, чтобы растягивать или сжимать любой график в двух направлениях.

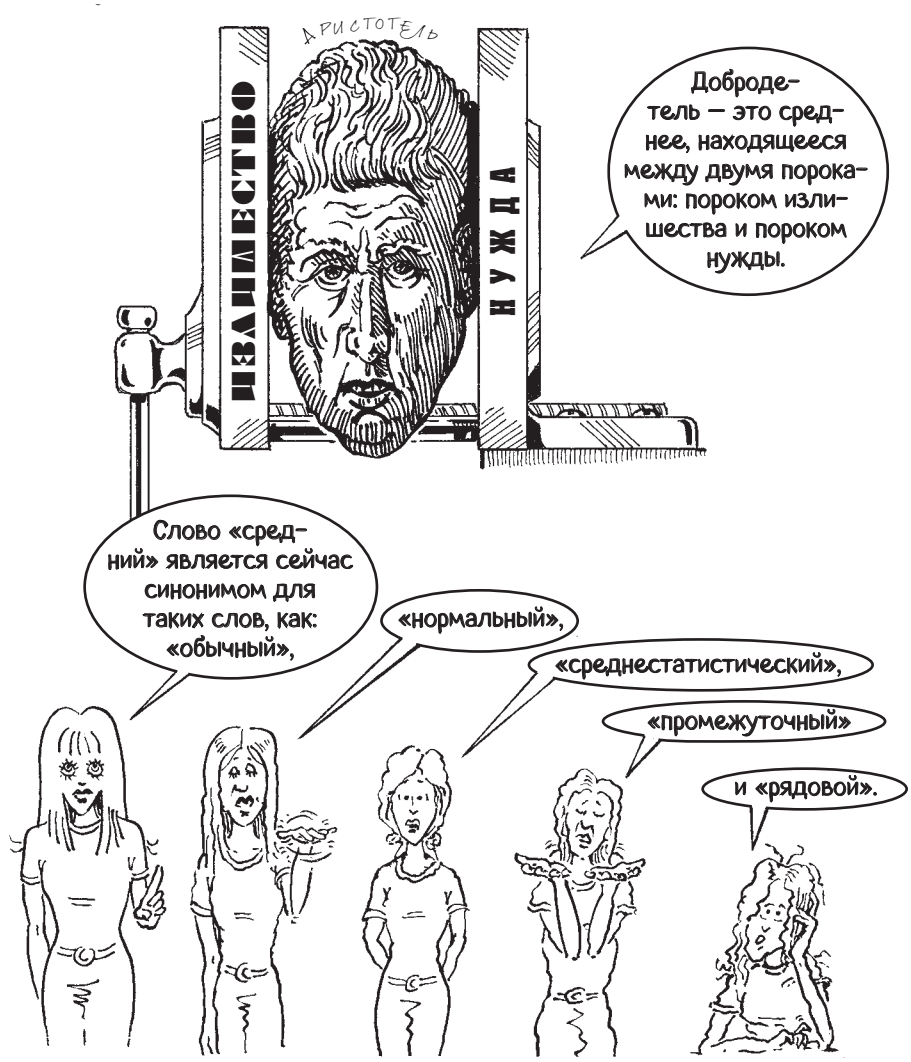


Эта бескомпромиссная вера в могущество нормальной кривой стала разделительной чертой между старой школой демографической статистики и новой, возникшей из математической статистики. Господство кривой нормального распределения было повсеместным, и к концу XIX века большинство статистиков признавали, что нет другой кривой для описания данных. Однако такой монолитный взгляд изменил в последней декаде столетия Пирсон.

Как суммировать данные?

Средние значения

Средние значения являются одним из основных инструментов демографической статистики и одним из старейших статистических понятий. Идея средних значений используется со времен античности. Аристотель писал о золотой середине, имея в виду «золотой» значит «хороший», находящийся между крайностями.

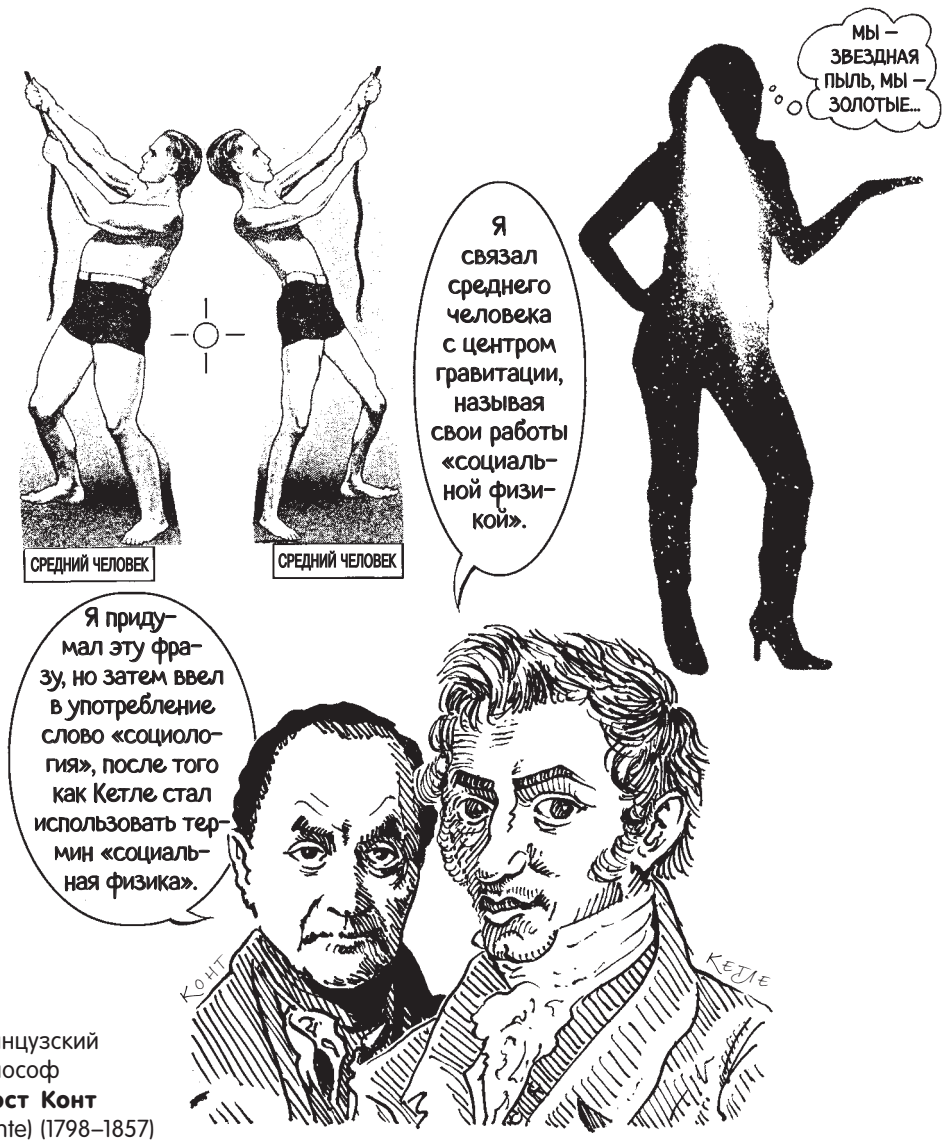


Но для статистиков существует три типа средних значений: среднее арифметическое, медиана и мода.

Кетле и среднее арифметическое

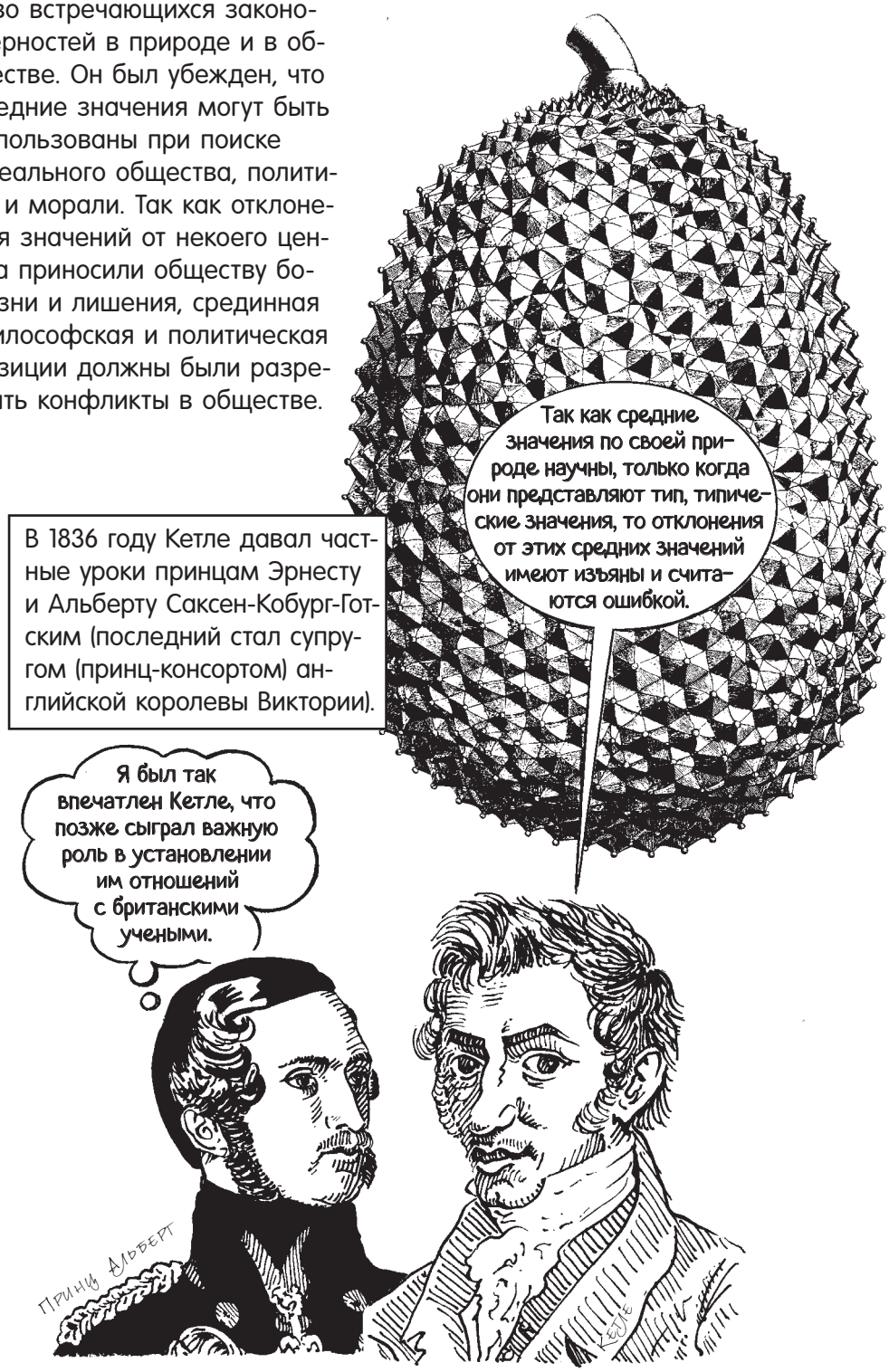
Такой метод был популяризован Кетле в 1830-х годах, когда было открыто, что астрономические законы ошибок можно применить к распределению таких характеристик человека, как рост и обхват талии. Это в свою очередь вызвало создание концепции l'homme moyen, или *среднего человека*.

Закономерности, которые Кетле нашел в человеке и в метеорах, были сопоставимы с законами физики. Он говорил об обществе так же, как астроном говорил о Вселенной.



Французский философ **Огюст Конт** (Comte) (1798–1857)

Кетле также подметил сходство встречающихся закономерностей в природе и в обществе. Он был убежден, что средние значения могут быть использованы при поиске идеального общества, политики и морали. Так как отклонения значений от некоего центра приносили обществу болезни и лишения, срединная философская и политическая позиции должны были разрешить конфликты в обществе.

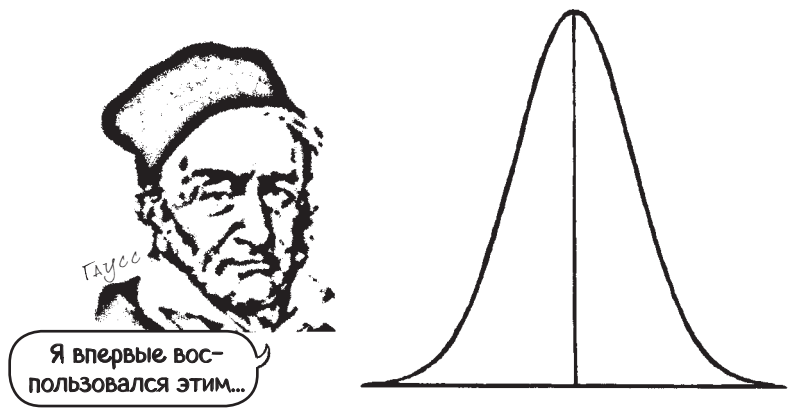


Среднее *арифметическое*

Среднее арифметическое — это то, что бóльшая часть людей привыкла считать собственно средним значением. Оно складывается из суммы всех значений набора данных (X), которое затем делится на общее число (N) случаев.

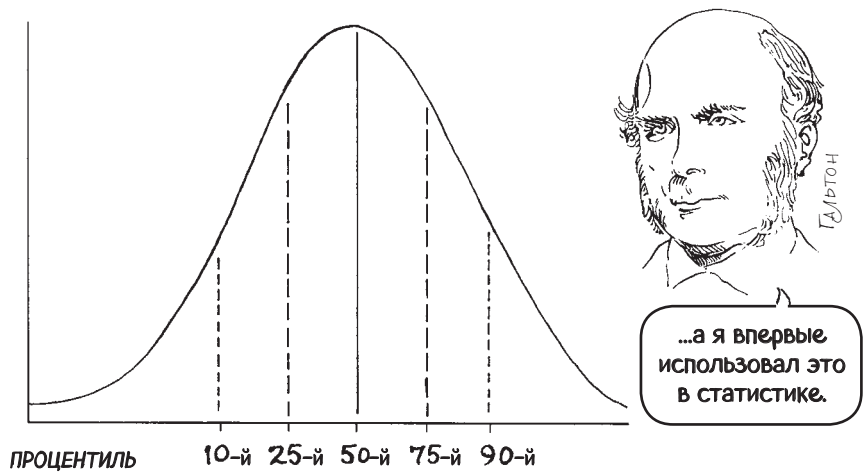
МЕДИАНА

Медиана — это точка, которая разделяет распределение на нижнюю и верхнюю половины таким образом, что количество значений в каждой из половин составляет 50% от общего.



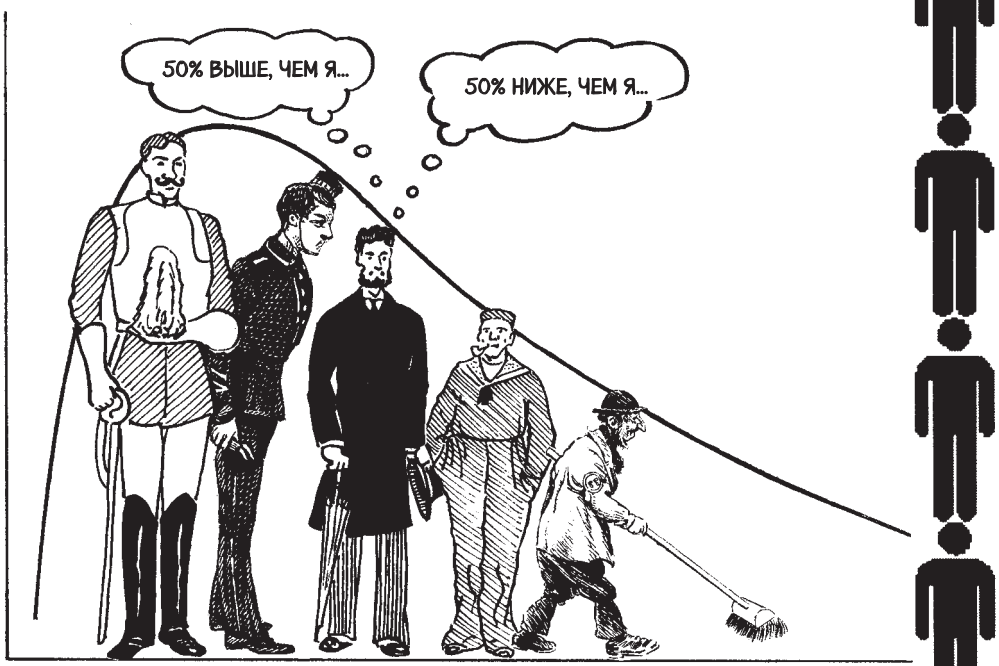
Фрэнсис Гальтон искал более быстрый способ определить середину, не сталкиваясь с трудностями подсчета значения среднего арифметического. Он ввел слово *процентиль*, которое обозначает точку, разделяющую распределение на нижние и верхние значения процентов.

Несмотря на то что в 1816 году Гаусс впервые воспользовался медианным значением, именно Гальтон ввел это понятие в статистику. В 1874 году он создал статистическую шкалу для того, чтобы найти медианное значение. Он использовал при этом 50-й процентиль как серединное значение в наборе данных, разделяющее эти данные строго на две равные половины.



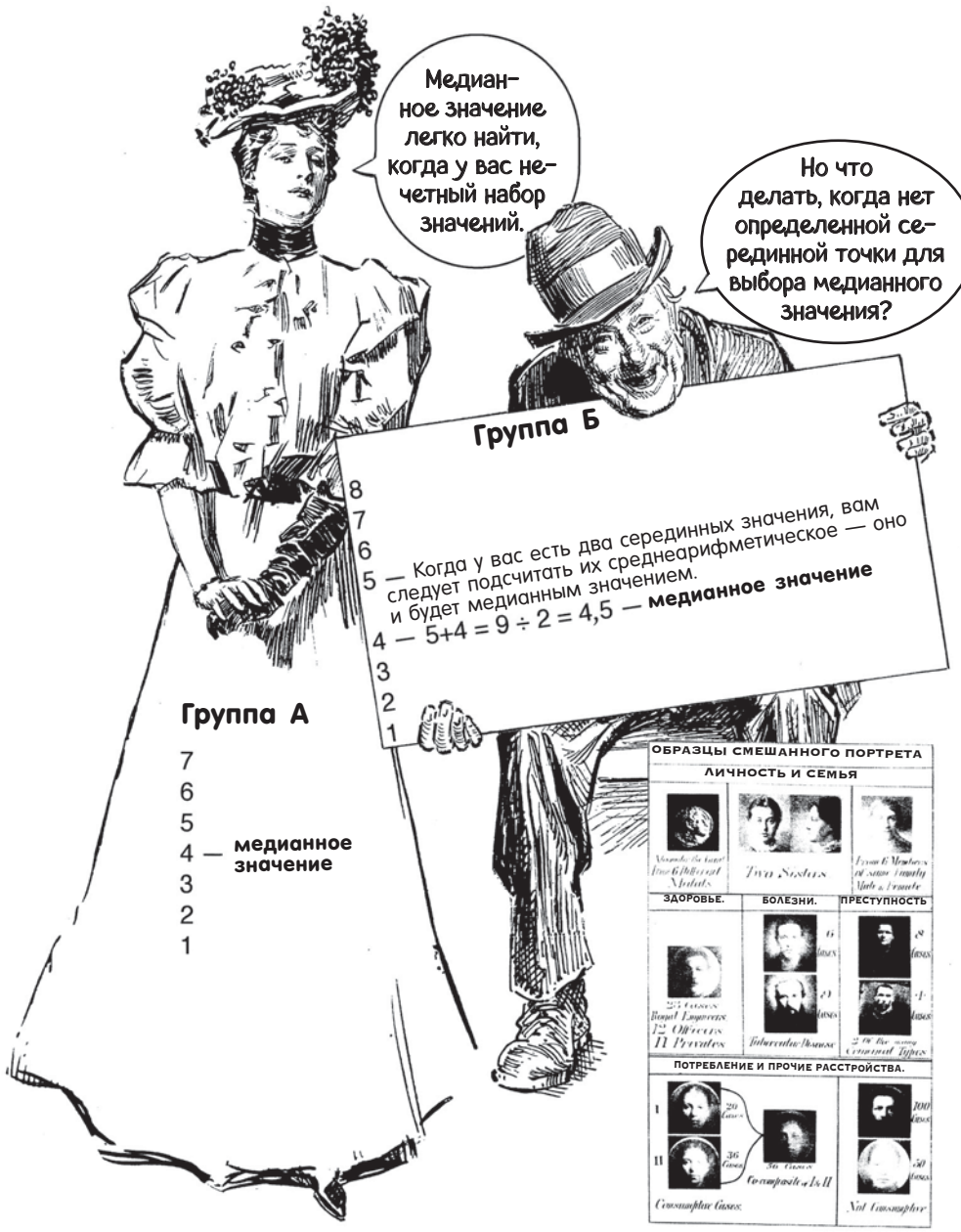
МЕДИАНА

Медианное значение относительно просто в использовании, и подсчитать его намного легче, чем среднеарифметическое. Когда Гальтон хотел измерить рост мужчин, он расставил в ряд 100 мужчин, от самого высокого до самого низкого, и выбрал того, кто стоял «как можно ближе к середине». Этот мужчина и представляет собой 50-й процентиль, или медианное значение.



МЕДИАННОЕ ЗНАЧЕНИЕ

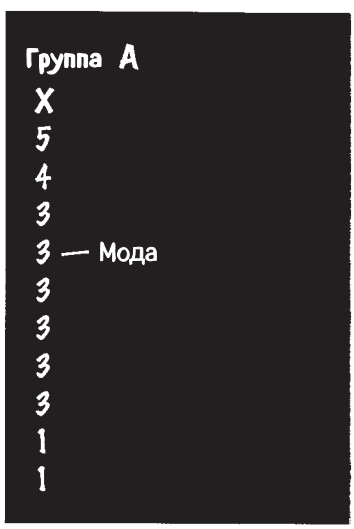
Как найти и подсчитать медианное значение?

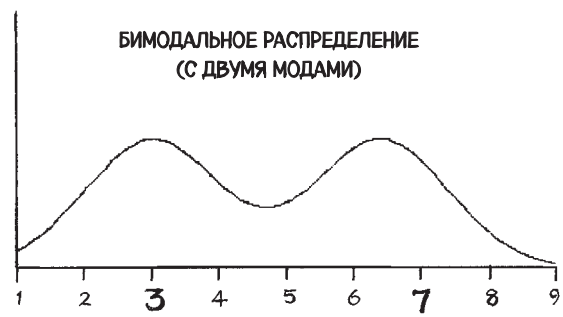


Гальтон даже изобрел способ создания (фотографирования) портрета среднего человека, накладывая друг на друга множество различных людей, которые сливались бы в одну картину — одно лицо. Гальтон назвал это комбинированной фотосъемкой.

Третий способ измерить среднее значение — мода — был придуман Карлом Пирсоном в 1894 году. Мода — это значение, которое встречается чаще других. Его часто используют в рекламе, когда нужно осуществить обращение к так называемой средней, или «репрезентативной» семье.

Мода — это точка наибольшей частости (частоты), она используется для того, чтобы рассмотреть типичные случаи. Моду вполне возможно (хотя и необязательно) сопоставлять с фактическим значением. «Средняя», или «модальная», семья, согласно исследованиям, может составлять 3,79 человека вместо 4-х.





В группе А есть значение, которое встречается 6 раз, и это значение — 3. Значит, 3 будет модой, однако в группе Б есть две моды: 7 и 3. Такой случай распределения статистической величины называется бимодальным.

Насколько важным является выбор статистического среднего?

Преимущество использования среднеарифметического состоит в том, что подсчет достаточно прост и включает весь набор данных в группе. Однако, если некоторые значения слишком велики или слишком малы, это исказит значение среднеарифметического.

В свою очередь, медианное значение не поддается влиянию крайних значений. Например, если вы захотите определить медианную зарплату среди такой группы — 40 000 фунтов, 60 000 фунтов, 120 000 фунтов, 160 000 фунтов, 820 000 фунтов, — медианным значением будет число 120 000 фунтов. Подобный метод поиска среднего будет полезным в подобной ситуации определения дохода, так как крайнее значение, 820 000 фунтов, искажает наблюдаемую картину и дает среднеарифметическое значение 240 000 фунтов, которое, как можно видеть, вовсе не является ничьей зарплатой.

Давайте рассмотрим все три способа измерения среднего для подсчета средней зарплаты в группе из 41 человека в компании.

X = один человек

Количество людей	Зарплата	
XX	4000 фунтов	
XXXXXX	6000 фунтов	
XXXXXXXX	10 000 фунтов	— Мода: значение, встречающееся наиболее часто
XXXX	18 000 фунтов	
X	24 000 фунтов	— Медианное значение: это середина, — выше находятся 20 человек и ниже находятся 20 человек
XXXX	30 000 фунтов	
XXX	36 000 фунтов	
XXXXX	40 000 фунтов	
XX	45 000 фунтов	
XXXX	50 000 фунтов	
X	70 000 фунтов	
X	200 000 фунтов	

Среднеарифметическое = **60 400** фунтов
Мода (встречается 8 раз) = **10 000** фунтов
Медианное значение = **24 000** фунтов

Как статистика может вводить в заблуждение

В этом примере все три цифровых подсчитанных значения для среднего отличались друг от друга. Мы сразу же можем заметить, что возможно намеренно ввести людей в заблуждение, выбирая то среднее, которое нам по каким-то причинам удобнее.

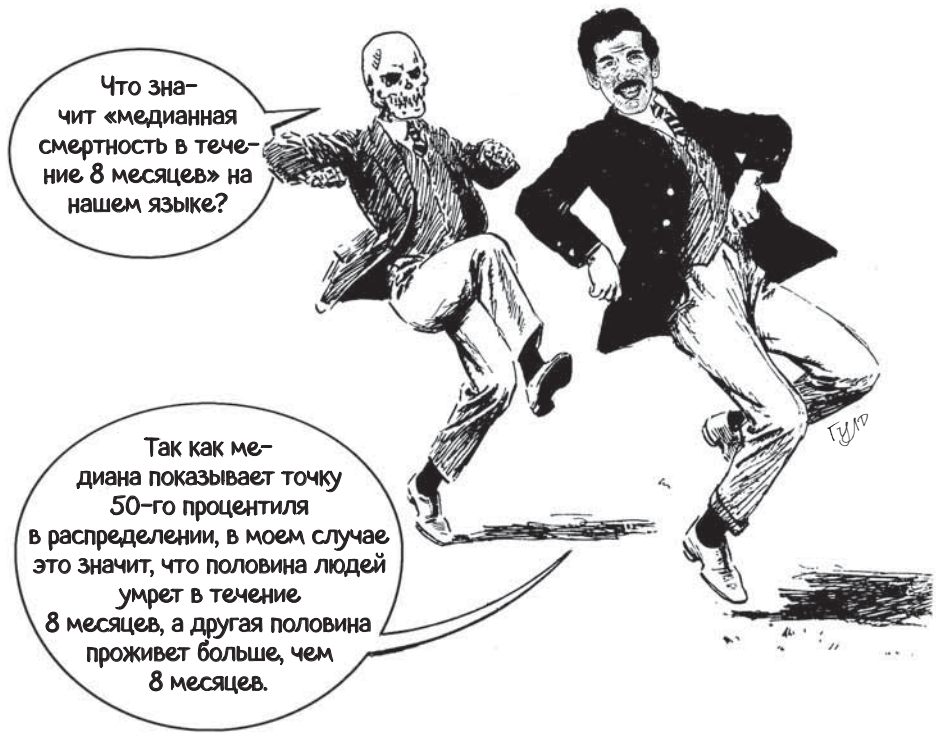
Пытливый журналист может заявить, что средняя (модальная) зарплата — 10 000 фунтов, и доказать, что половина сотрудников в компании получают меньше среднего дохода по стране в целом.

Медианное значение (24 000 фунтов) здесь, пожалуй, наиболее репрезентативно, хотя результат мог бы быть еще более реалистичным, если бы зарплата босса (200 000 фунтов) не была включена. Она представляет собой крайнее значение по сравнению с остальными. Статистики называют такие крайние значения выпадающими, так как они находятся далеко на краях распределения.

Частотное распределение 37 человек из нашего примера (гистограмма)

Единственным способом рассуждать о средних значениях является учет всего объема информации, в особенности вариаций около среднеарифметических значений. Зачастую это наиболее реалистичный способ обнаружить характерную информацию относительно индивида.

Это был полезный урок, который палеонтолог и биолог-эволюционист Стивен Джей Гулд (Gould) (1941–2002) усвоил вскоре после того, как у него нашли в 1982 году мезотелиому (редкий и тяжелый тип рака, обычно вызываемый работой с асбестом). Его знание статистики помогло ему понять, что он не должен уподобляться простому статистику, который верит в медианную смертность от рака в течение 8 месяцев, — согласно тогдашним прогнозам медиков.



Важным практическим инструментом для работы со статистическими данными является **частотное распределение** (с. 35). Гулд понимал, что этот график не означает его неминуемую смерть в течение 8 месяцев. Напротив, график можно проинтерпретировать так, что сам Гулд мог с легкостью оказаться справа от медианного значения — среди пациентов, которые прожили больше 8 месяцев.

Гулд рассчитывал на то, что большинство людей, незнакомых со статистикой, поймут «медианную смертность в течение 8 месяцев» как «Я умру в ближайшие 8 месяцев».

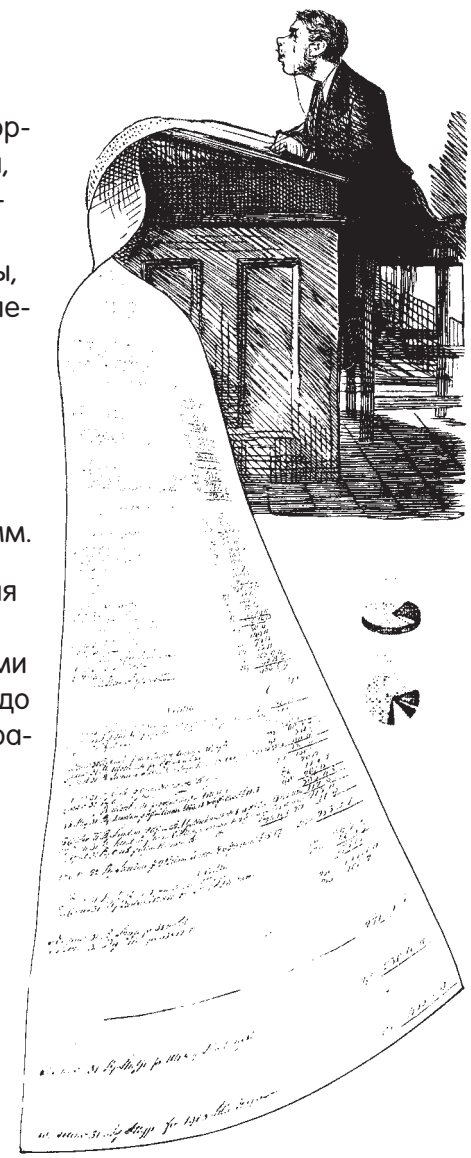
Будучи биологом-эволюционистом, Гулд знал, что нужно исследовать изменчивость в качестве основы реальности и стараться избегать средних значений, которые в конце концов являются лишь абстрактной мерой, неприменимой к отдельному человеку или неуместной в индивидуальных случаях.

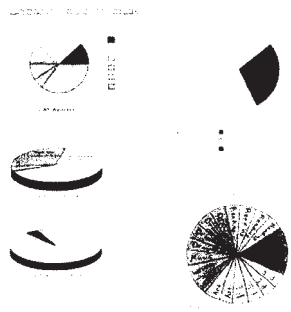
В свете прозорливости Гулда один из колумнистов *«Санди Таймс»* проницательно предположил, что «Статистика — это приговоренный лучший друг человека». Стивен Джей Гулд умер в 2002 году, через два десятилетия после первичного диагноза.

Способы управления данными

Викторианцы были одними из первых, кто использовал статистику для изучения массовых явлений. Колоссальное количество данных было собрано государственными агентствами, частными организациями и различными личностями, заинтересованными в таких общественных явлениях, как бедность, болезни и суицид. Существуют основные способы, которыми они пользовались для управления данными.

1. Группировка (составление таблиц) — простая запись данных в длинные столбцы цифр.
2. Создание круговых и прочих диаграмм.
3. Сужение набора данных для создания выборок меньшего размера. Например, когда Гальтон работал с крупными выборками, он часто сужал выборку до 100 человек, для наглядной демонстрации процентных отношений.





Но так как диаграммы и таблицы не имели стандартов построения, обобщения или сравнения с другими наборами данных были невозможны. Хотя викторианцы и использовали средние значения для формирования выводов из анализа своих данных, все-таки их статистический инструментарий не мог передать всю сложность этих данных, присущую моделям, которые имели дело уже со статистической изменчивостью (variation).

Унифицированные частотные распределения

Пирсон обнаружил, что имеются и другие способы организации и упорядочения громоздких данных. Он разработал систематический способ для работы с очень большими наборами данных, создав **унифицированное частотное распределение**. Оно позволило сравнивать и обобщать те данные, с которыми раньше было невозможно работать.

Выборки или генеральные совокупности?

В 1892 году близкий друг Пирсона, последователь Дарвина и зоолог В. Ф. Р. Велдон (Weldon) (1860–1906) ввел в обращение термин «выборка» для обозначения групп наблюдений за морскими организмами, хотя и полагал, что размер его выборок достаточно большой. Пирсон использовал термин «генеральная совокупность» четырьмя годами позже, заменив термин «нормальная группа» и поставив *генеральную совокупность* в один ряд с *выборкой* в 1903 году.

Генеральная совокупность — это технический термин, обозначающий целую группу организмов или объектов, таких как розы или тигры, на которые распространяются результаты. Генеральная совокупность представляет все возможные варианты наблюдений определенного типа, в то время как *выборка* — это ограниченное число наблюдений из генеральной совокупности. Наилучшим примером использования целой генеральной совокупности является перепись (например, населения), проводящаяся каждые десять лет.

Генеральная совокупность

Выборка

В большинстве исследований генеральная совокупность, в которой заинтересован аналитик, слишком велика для измерения всех ее элементов (все студенты Англии, все голосующие в Великобритании, все машины «Форд» и т. д.). Ученые-статистики обычно ограничивают свой анализ генеральной совокупности какой-либо небольшой группой наблюдений внутри генеральной совокупности, которая называется выборкой.

Статистики используют несколько методик выборочного исследования: **случайную, систематическую, побочную, целевую и расслоенную** (stratified).

Случайная выборка

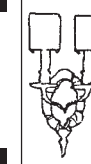
Этот тип аналогичен вытягиванию нескольких ярлыков с именами людей из шляпы, в которую эти ярлыки (в очень большом количестве) были прежде положены. Каждый элемент в такой генеральной совокупности независим от других и *обладает одинаковой с ними вероятностью* попасть в выборку. И хотя этот тип выборки наиболее приемлемый, для него необходимо иметь полный список всех элементов генеральной совокупности, который не всегда можно составить. Таблица случайных чисел в статистических сборниках или аналоги, создаваемые компьютерами и некоторыми телефонными системами, используются при рассмотренном типе.

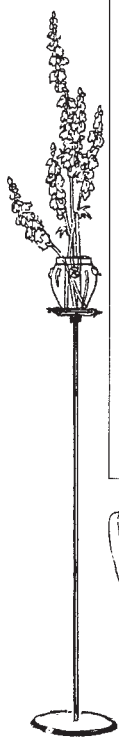
Систематическая выборка

Для этого типа также требуется полный список элементов генеральной совокупности, однако здесь он разделен на блоки, в каждом из которых выбирается каждый n-й элемент из списка (например, выбирается каждый 10-й элемент из отсортированного по алфавиту списка).

Побочная выборка

Этот тип является наиболее доступным, так как здесь выборка составляется из наиболее удобного и подходящего набора элементов. Однако такой тип выборки является самым недостоверным из всех.



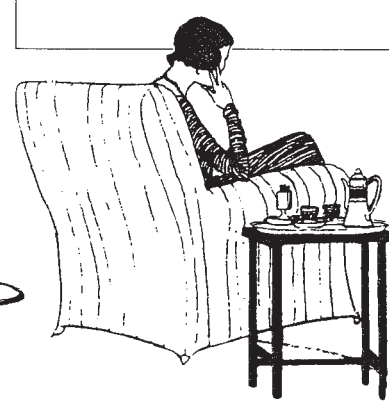


Целевая выборка

В этом типе исследователь сам выбирает элементы для своей выборки, потому что он или она считает их наиболее репрезентативными.

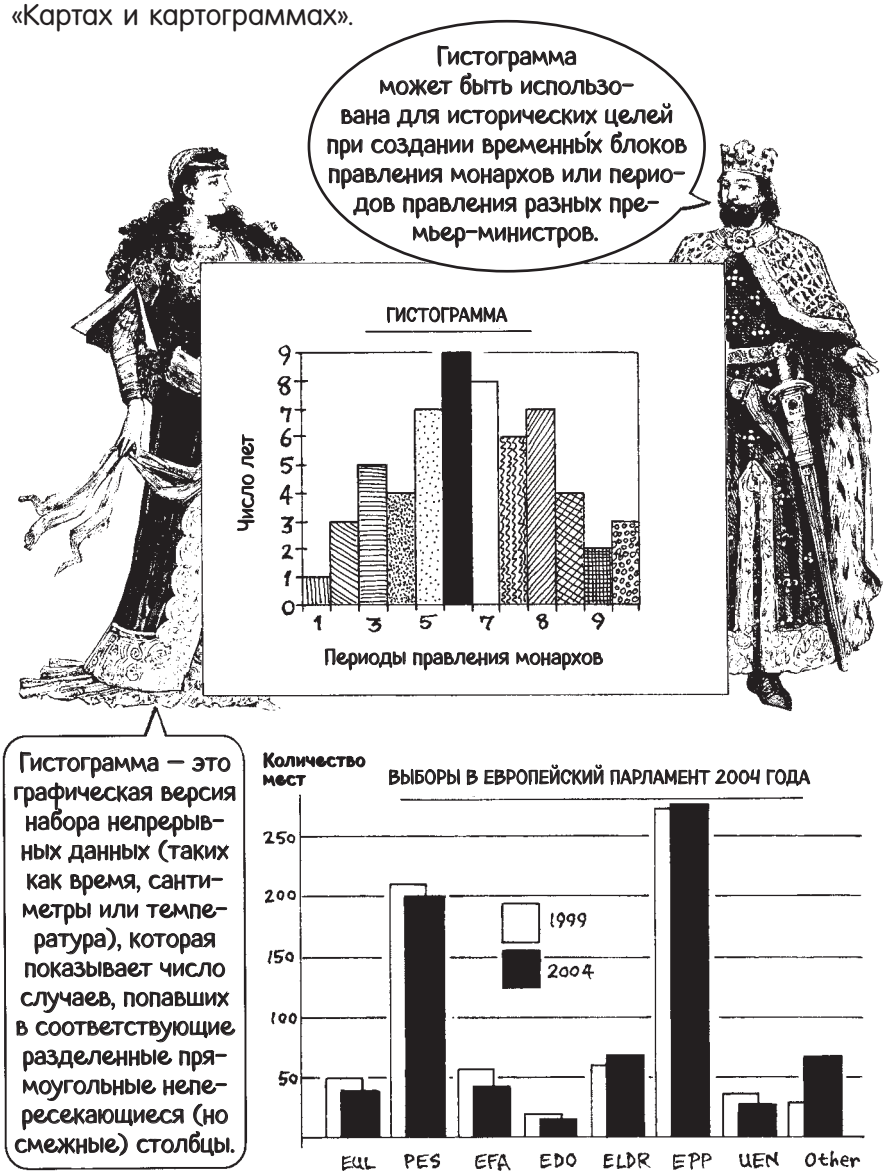
Расслоенная выборка

Используя расслоенную выборку, исследователь выбирает определенную характеристику, которую он или она считает важной для исследования, а затем разделяет выборку на непересекающиеся группы или слои, страты (например, возрастные, гендерные, географические или политические). Этот тип можно использовать в сочетании с одной из предыдущих четырех выборок.



Гистограмма

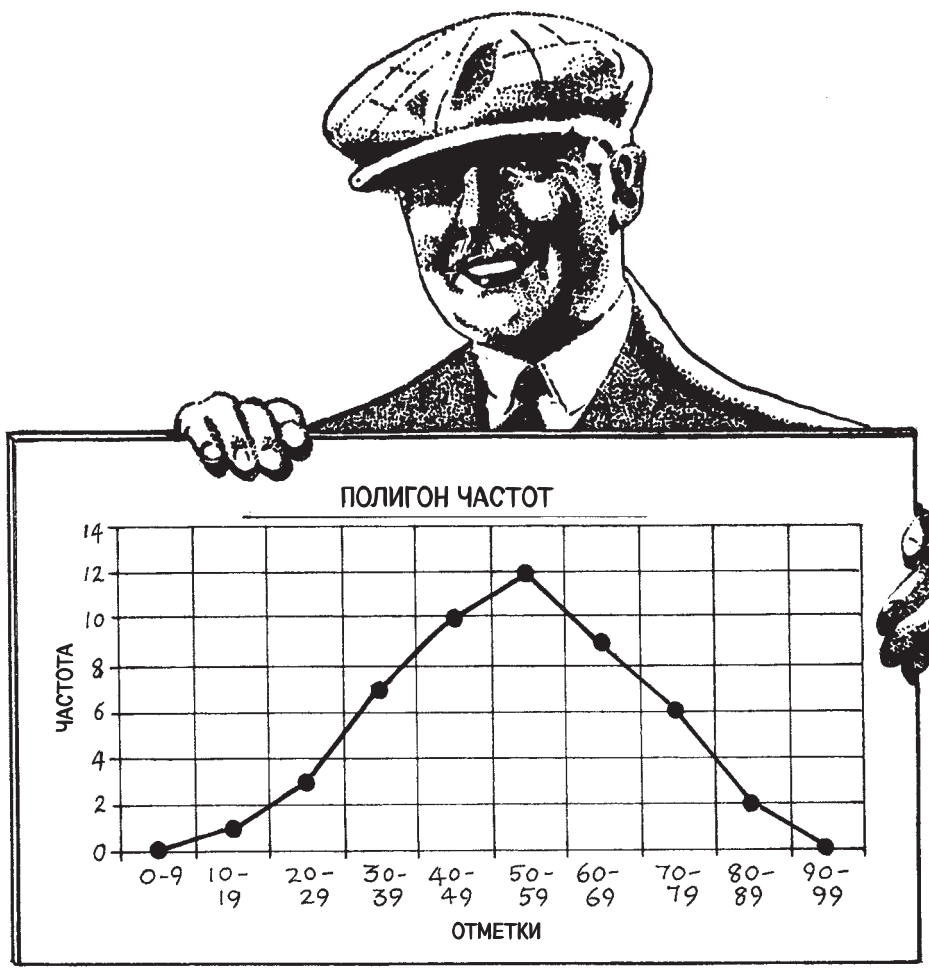
Пирсон ввел в употребление гистограмму 18 ноября 1891 года. Он придумал этот термин для обозначения «временно́й диаграммы» на своей лекции о «Картах и картограммах».



График, который внешне похож на гистограмму, называется столбчатым графиком. В нем есть зазоры между столбцами, и его построения используются дискретные данные (такие как пол, политическая принадлежность). К графикам часто прибегают, чтобы помочь людям взглянуть на проблему визуальными средствами.

Другой способ представить набор непрерывных данных заключается в использовании полигона частот. Полигон частот — это линейный график, который состоит из срединных точек каждого столбца (взятого из гистограммы) и соединенных прямой линией.

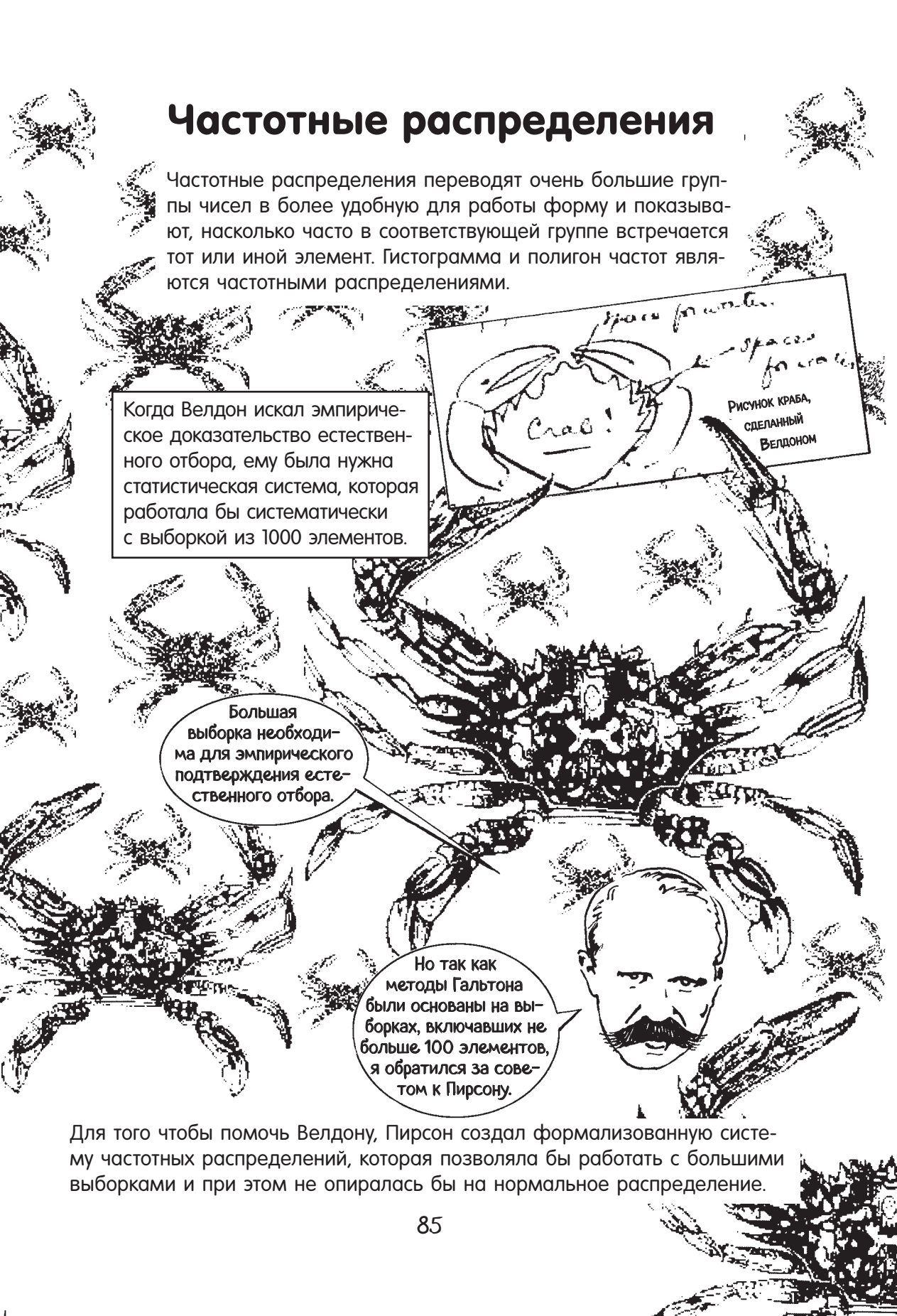
Процесс нанесения данных на картинку полигона частот — это наиболее простой тип вычерчивания эмпирической кривой по точкам, который заключается в соединении двух точек прямой линией (или более сложной кривой) для создания различных форм статистической зависимости.



Следующим шагом для Пирсона была демонстрация студентам того, как можно скомпоновать частотные распределения на случай больших объемов непрерывных данных и как сконструировать такие распределения.

Частотные распределения

Частотные распределения переводят очень большие группы чисел в более удобную для работы форму и показывают, насколько часто в соответствующей группе встречается тот или иной элемент. Гистограмма и полигон частот являются частотными распределениями.



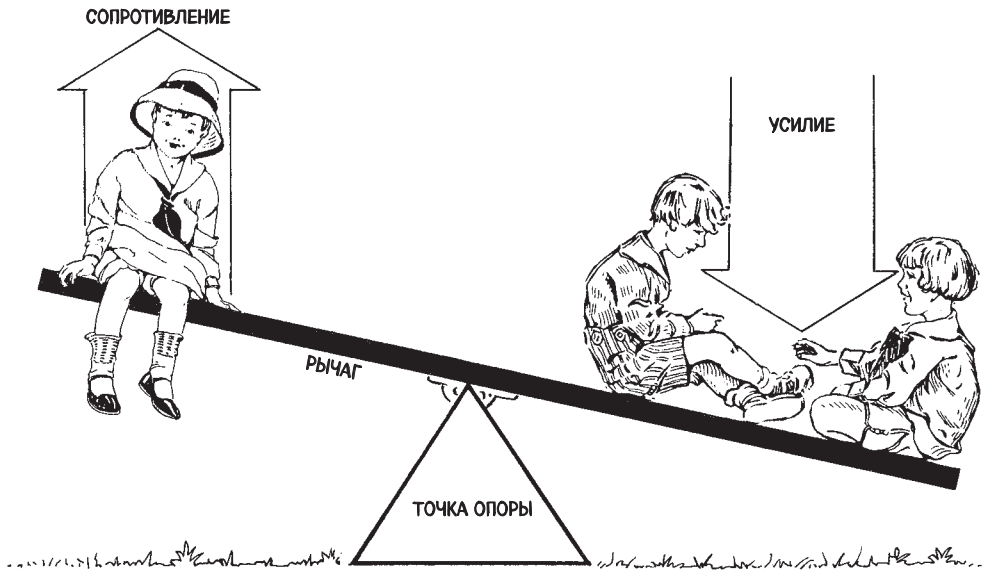
Для того чтобы помочь Велдону, Пирсон создал формализованную систему частотных распределений, которая позволяла бы работать с большими выборками и при этом не опиралась бы на нормальное распределение.

Метод моментов

Как определить и описать форму эмпирического распределения?

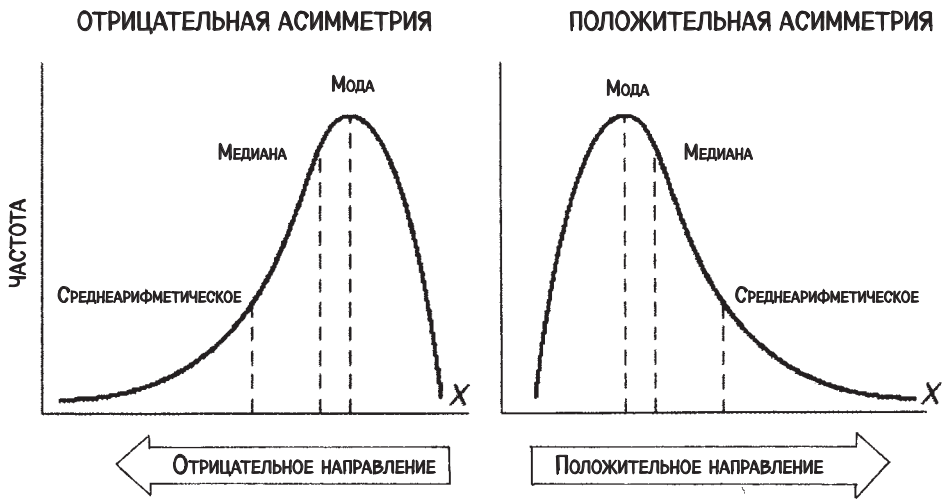
Пирсон начал разрабатывать свою статистическую систему в 1892 году, основываясь на **методе моментов**. Термин «момент» пришел из механики: он измеряет силу, приложенную к точке вращения, например к точке опоры рычага. В статистике моменты — это средние значения. Вычислительные процедуры в отношении моментов аналогичны поиску среднеарифметического. Пирсон заменил механическую силу функцией кривой распределения частот (такой, которая показывала бы процентное распределение внутри заданного интервала группировки).

Будучи заядлым любителем графических представлений, Пирсон объяснял метод моментов своим студентам, используя примеры из механики. Для вычисления среднеарифметического он нашел точку, в которой рычаг балансирует на точке опоры. Среднеарифметическое является «точкой баланса» этого рычага и аналогично центру гравитации (или масс) в механике.



Если приложить силу к такому рычагу, первый момент будет называться «моментом силы». Вычисления производятся для того, чтобы определить первый момент и найти среднеарифметическое. Пирсон продолжал использовать эту процедуру и со следующими тремя моментами. Используя одни и те же данные при поиске среднеарифметического, он возводил в квадрат полученные значения для того, чтобы найти квадрат *среднеквадратического отклонения* (см. с. 99–102).

Для того чтобы измерить асимметрию распределения, Пирсон возводил в третью степень эти средние значения и вычислял третий момент. Когда распределение асимметрично, среднее располагается ближе к хвосту распределения.

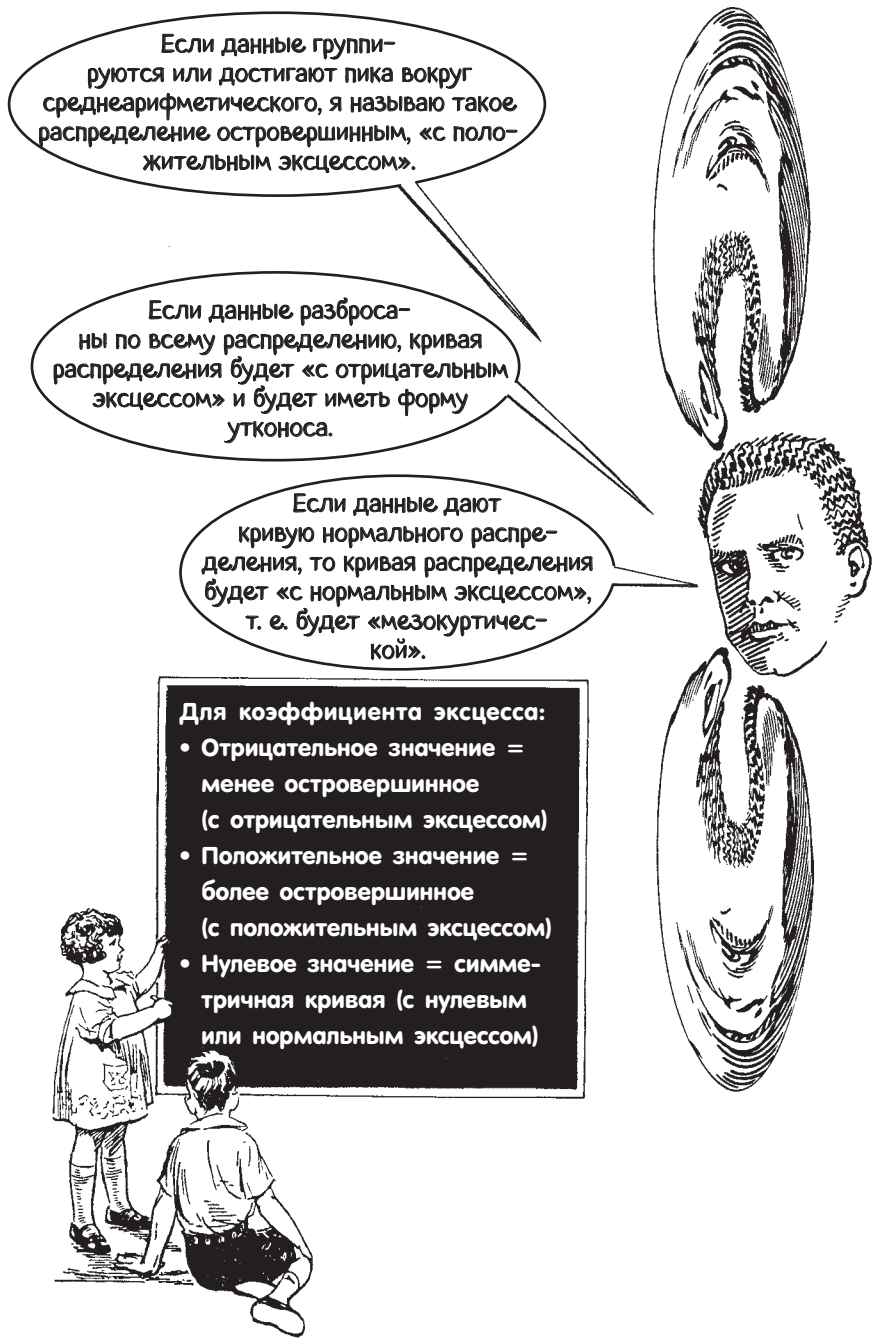


Значение асимметрии:
Если значение равно 0, значит, распределение симметрично.
Если значение отрицательно, значит, присутствует отрицательная асимметрия.
Если значение положительно, значит, присутствует положительная асимметрия.

Первый коэффициент асимметрии Пирсона позволил ему вычислить асимметрию, подсчитывая разницу между среднеарифметическим и модой, разделенную на среднеквадратическое отклонение.

$$\textbf{Асимметрия} = \frac{\textbf{(среднеарифметическое – мода)}}{\textbf{среднеквадратическое отклонение}}$$

Для вычисления четвертого момента Пирсон возводил средние значения в четвертую степень. Это показывало, насколько плоским или островершинным было распределение. Пирсон придумал слово *kurtosis* [рус. «коэффициент эксцесса»] для обозначения этого момента (от греческого слова, обозначающего «вздутость»). Соответственно, есть три варианта значений этого показателя.



Один из студентов Пирсона **Уильям Сили Госсет** (Gosset) (1876–1937), который известен под псевдонимом «Стьюдент», использовал иллюстрацию утконоса для изображения кривой с отрицательным эксцессом и двух кенгуру с длинными хвостами для кривой с положительным эксцессом.

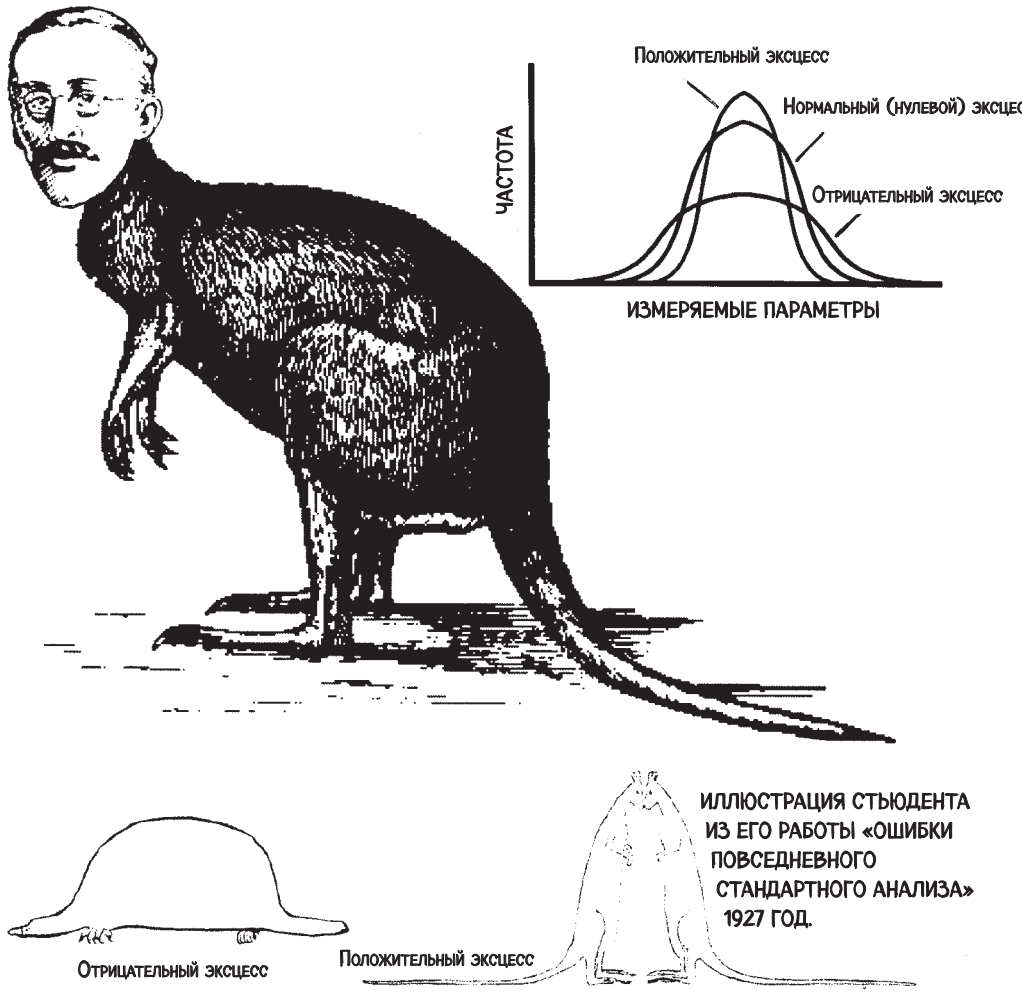
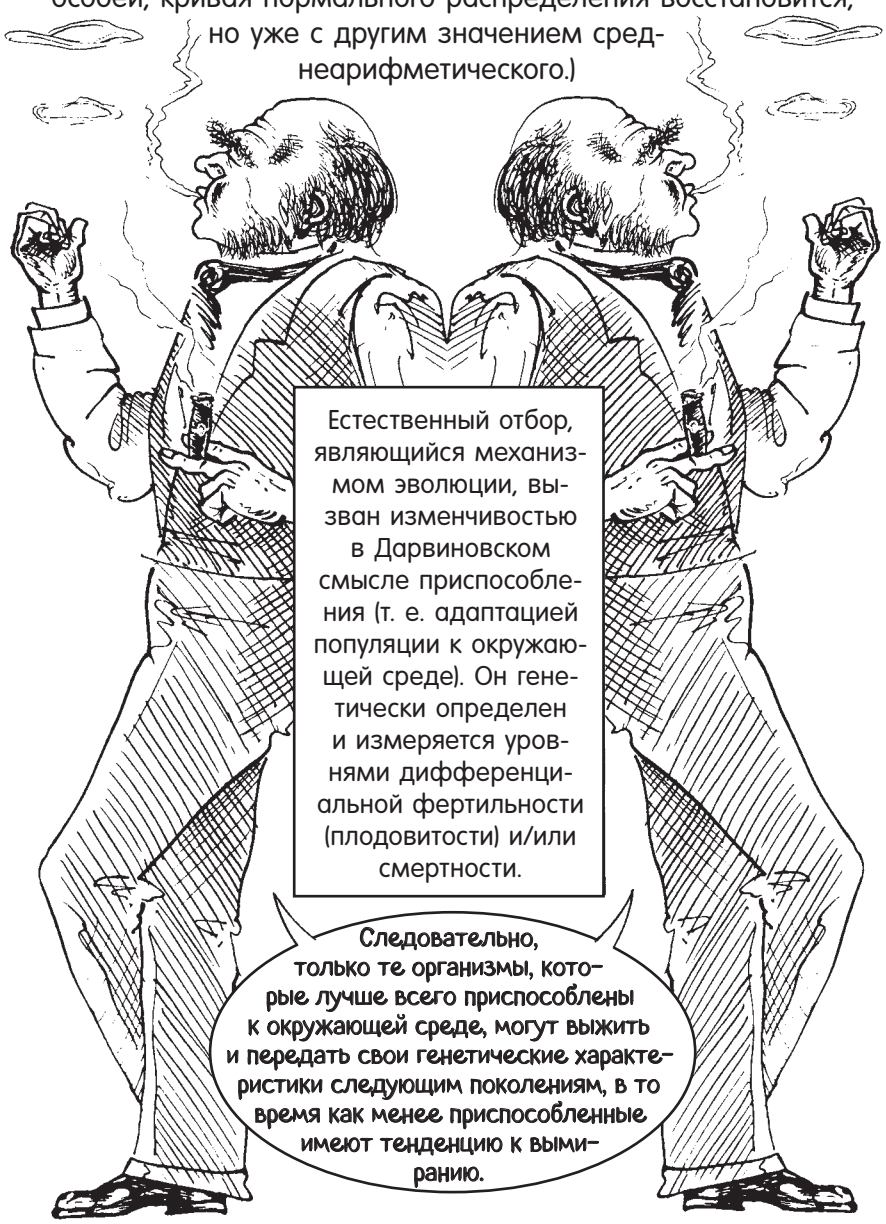


ИЛЛЮСТРАЦИЯ СТЬЮДЕНТА ИЗ ЕГО РАБОТЫ «ОШИБКИ ПОВСЕДНЕВНОГО СТАНДАРТНОГО АНАЛИЗА» 1927 ГОД.

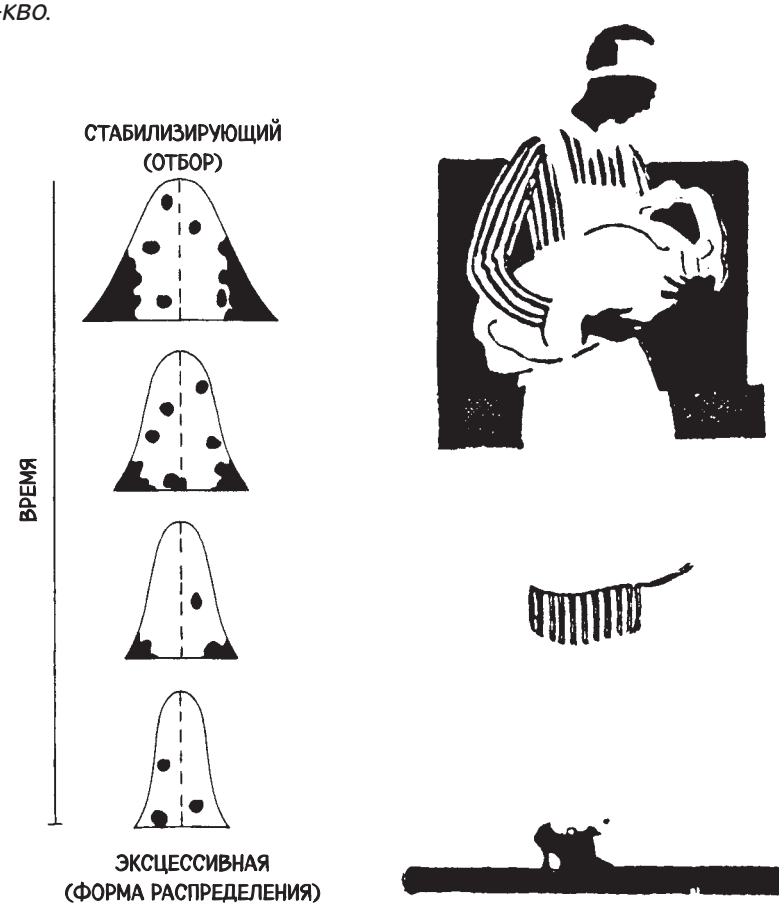
Используя метод моментов, Пирсон установил четыре параметра для статистической обработки и построения кривых. Эти параметры показывали, соответственно, как 1) *сгруппированы* данные (среднеарифметическое), 2) каков у них *разброс* (среднеквадратическое отклонение), наблюдается ли 3) потеря *симметрии* (асимметрия) и 4) какова форма распределения — *островершинная* или *плоская* (коэффициент эксцесса). Эти четыре параметра описывали основные характеристики любого распределения: система была экономной и элегантной. Эти статистические инструменты необходимы для интерпретации любого набора статистических данных, какова бы ни была форма их распределения.

Естественный отбор: изменяющиеся формы Дарвиновских распределений

Дарвин понимал, что форма частотного распределения до появления естественного отбора была бы «симметричной относительно среднеарифметического» (т. е. данные имели бы нормальное распределение) и что после начала работы механизма естественного отбора распределение утратит свою симметричную, колоколообразную форму. (Однако затем, по мере размножения и воспроизводства потомства особей, кривая нормального распределения восстановится, но уже с другим значением среднеарифметического.)



Если форма распределения островершинная или сплюснутая (эксцессивная, по терминологии Пирсона), то можно предполагать *стабилизирующий отбор*, который означает поиск поддержания баланса, или состояния *статус-кво*.

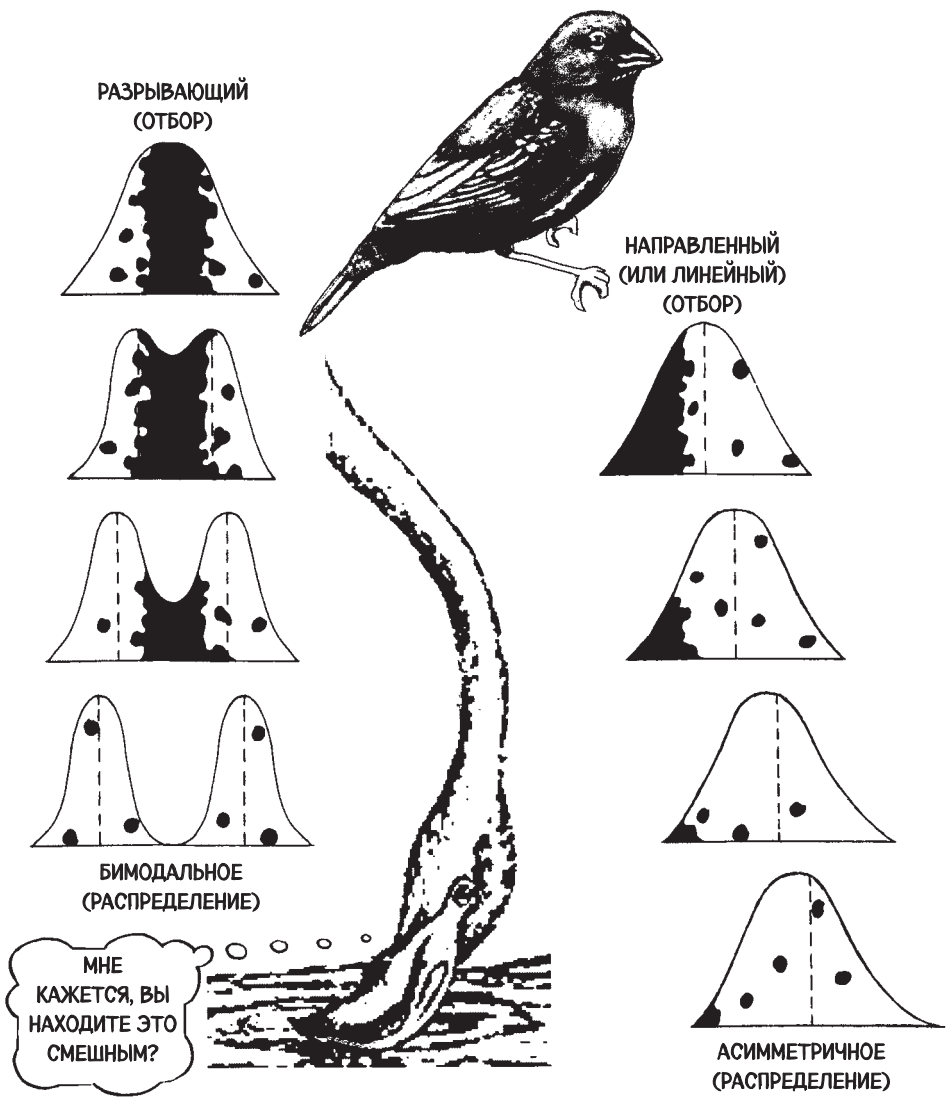


Распределение сверху — это нормальное распределение до естественного отбора. Черные области обозначают места давления отбора за определенное время, до тех пор пока форма распределения не изменится на ту, которая представлена внизу.

Давление отбора (или селективное давление) — это любое из ряда вон выходящее явление, которое заставляет менять поведение и приспособленность живых организмов внутри рассматриваемой окружающей среды. Оно представляет собой движущую силу эволюции и естественного отбора.

Масса тела детей при рождении находится под определенным влиянием стабилизирующего отбора. Детская смертность меньше всего при средней массе тела и максимально высока при слишком низкой или слишком высокой массе новорожденного тела.

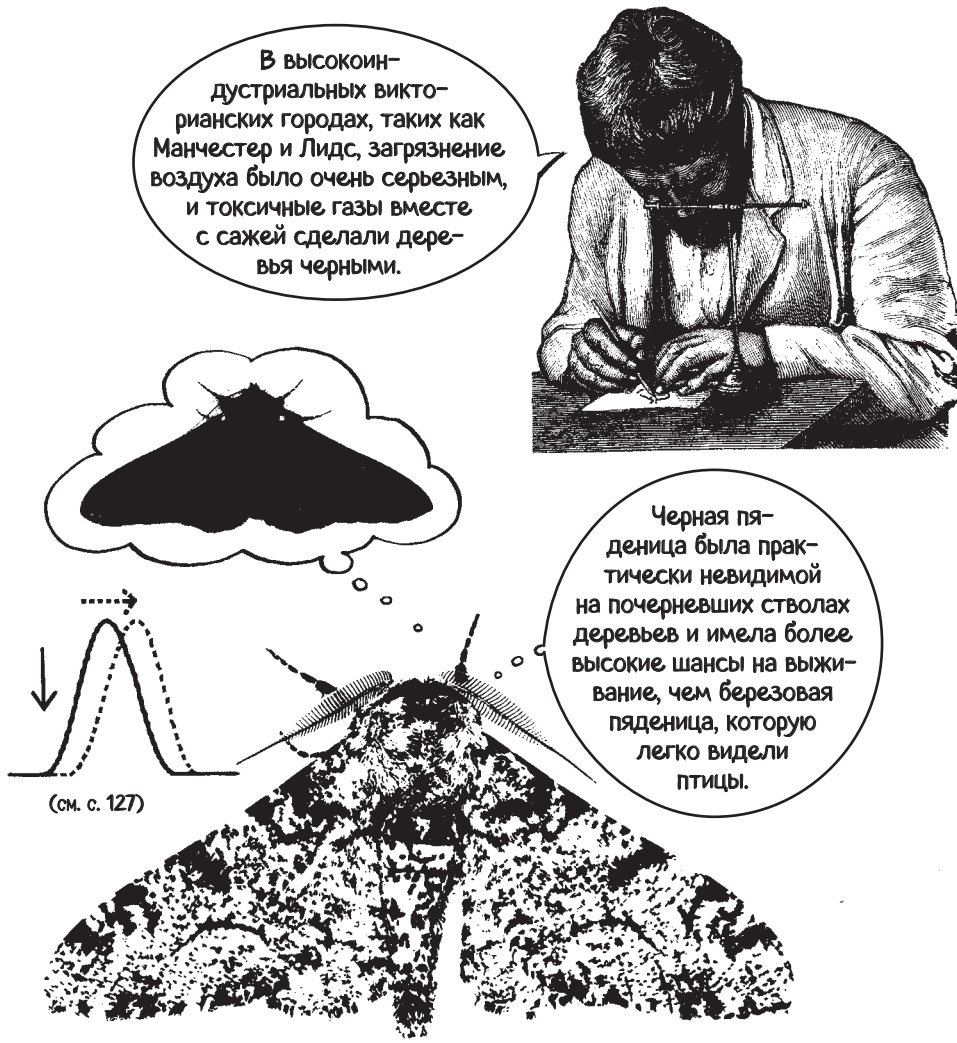
Бимодальное распределение обозначает *разрывающий отбор* (или *разрушительную селекцию*), при котором уничтожается середина распределения и остаются только его края. Разрывающий отбор был обнаружен у чернобрюхих астрильдов *(Pyrenestes ostrinus)*, которые обитают в Западной Африке. Птенцы с маленькими клювами ели мягкие, небольшие семена, а птенцы с крупными клювами ели большие, твердые семена.



Если распределение в одном из направлений становится асимметричным, это свидетельствует о *направленном отборе*, который происходит, когда популяция находит обстоятельства на одном конце распределения более привлекательными, чем на другом.

Пяденица березовая

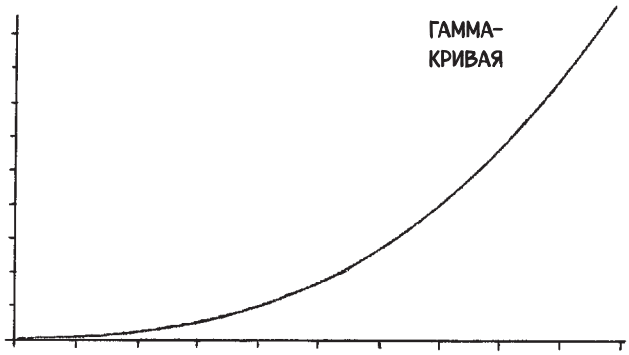
Один из наиболее известных примеров *направленного отбора* встречается у березовых пядениц *(Biston betularia),* которые обитали в огромном количестве в предындустриальной викторианской Англии. Хотя угольно-черный мутант был обнаружен в 1849 году, они все равно редко встречались уже в то время.

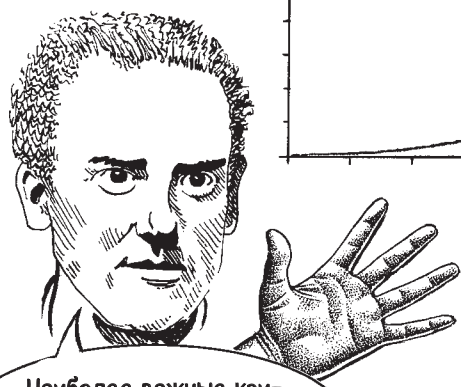


В течение века популяция угольно-черных пядениц возросла на 90% на индустриальном Севере. В то время как исконные березовые пяденицы, имеющие до индустриализации нормальное распределение, сразу после загрязнения среды их обитания изменили свое распределение: кривая нормального распределения сместилась в правую часть и стала асимметричной.

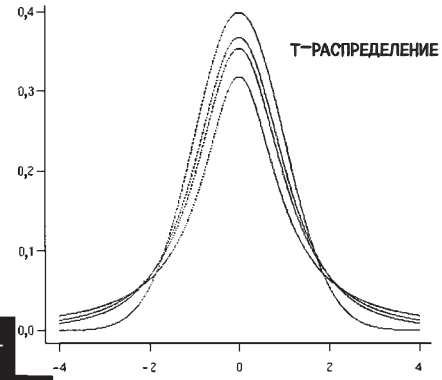
Пирсоновское семейство кривых

Используя метод моментов, Пирсон также создал целый ряд теоретических кривых с различной градуировкой, которые затем могли быть наложены на эмпирическую кривую для определения того, какая кривая подходит наилучшим образом. Эти кривые относятся к «Пирсоновскому семейству кривых».





Наиболее важные кривые, которые остаются основой теоретической статистики, включают в себя:

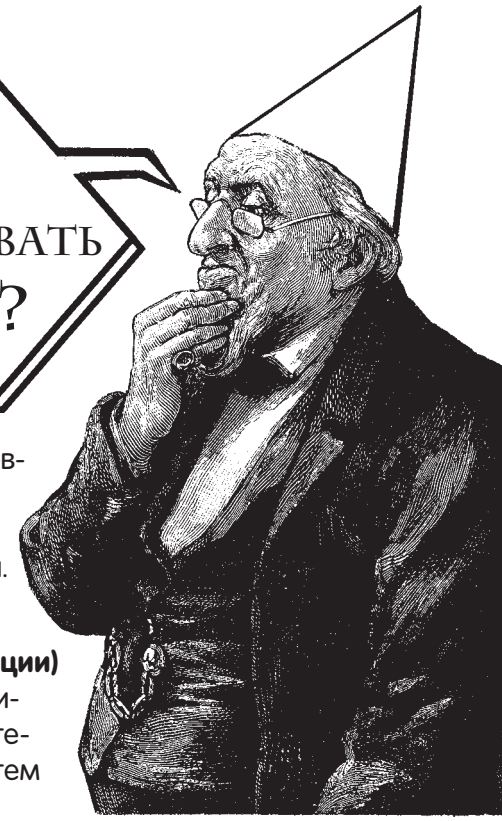


III тип: *гамма-кривая, которую он использовал в попытке найти точное распределение хи-квадрат (мы рассмотрим это распределение позднее).* ***IV тип:*** *семейство асимметрических кривых (созданных для данных Велдона).* ***V тип:*** *кривая нормального распределения.* ***VII тип:*** *ныне известная как распределение Стьюдента для получения t-статистик в тестах на проверку гипотез (тема будет затронута позднее).*

Открытие Пирсоном этого семейства кривых сделало многое для того, чтобы развенчать почти религиозное убеждение в том, что нормальное распределение является математической моделью изменчивости биологических, физических и социальных явлений.

Черчилль Эйзенхарт (Eisenhart)
(1913–1994)

КАК ИНТЕРПРЕТИРОВАТЬ ДАННЫЕ?



Статистика начала исследовать основные закономерности и типы изменчивости и любые ярко выраженные отклонения от этих закономерностей.

Статистические измерения изменчивости (вариации)

Измерение изменчивости (или вариации) — это ключевой элемент математической статистики и вместе с тем поворотный момент в ее развитии. Гальтон придумал первый способ измерения в 1875 году, когда ввел понятие «полуинтерквартильный размах» (полурасстояние между квартилями). Он выражался так: $\frac{Q_3 - Q_1}{2}$,

где Q, квартиль — это точка на кривой распределения.

1% – 25%	26% – 50%	51% – 75%	76% – 100%
Q_1	Q_2	Q_3	Q_4
Первый квартиль	**Второй квартиль**	**Третий квартиль**	**Четвертый квартиль**

Гальтон

Полуинтерквартильный размах

Подобно Гальтоновой медиане, этот метод был легким и быстрым в использовании. На полуинтерквартильный размах не влияли выпадающие значения.

2 3 4 6 **6** 8 9 11 **12** 14 14 15 **17** 18 19 21 **82**

Q_1 Q_2 Q_3 **выпадающее значение**

Здесь полуинтерквартильный $= \frac{17 - 6}{2} = \frac{11}{2} = 5{,}5$

Интерквартильный размах

Это более широко используемый метод, который измеряет разброс 50% середины (или медиану) организованного набора данных. В следующем примере:

1 1 3 4 **4** 5 5 6 **6** 7 7 8 **8** 9 9 9 10
Q_1 **Q_2** **Q_3**

интерквартильный размах = Q_3–Q_1 или 8–4=4. Следовательно, медиана (Q_2) равняется 6, и разброс происходит в радиусе 4. Эта техника оставалась легким и быстрым способом ручного подсчета приблизительной оценки значений изменчивости вплоть до появления статистического программного обеспечения для ПК в конце 1970-х годов.

Подобно полуинтерквартильному размаху, на интерквартильный размах не оказывают влияние выпадающие значения.

2 3 4 6 **6** 8 9 11 **12** 14 14 15 **17** 18 19 21 **82**
Q_1 **Q_2** **Q_3** **выпадающее значение**

Интерквартильный размах = Q_3–Q_1 или 17–6=11. Вокруг медианы, которая равняется 12, происходит разброс в радиусе 11.

РАЗМАХ

В 1892 году в своих первых лекциях по статистике в Грешем-колледже Пирсон ввел понятие **размаха**, которое является простейшим способом измерить изменчивость. Размах измеряет расстояние между наибольшим и наименьшим значениями из определенного набора данных наблюдений и дает представление о разбросе данных.

В данном примере — 4, 7, 12, 25, 34 — размах будет равен 34–4=30

Размах часто используется при изложении данных для обычных граждан, например размах зарплат, возрастов и температур.

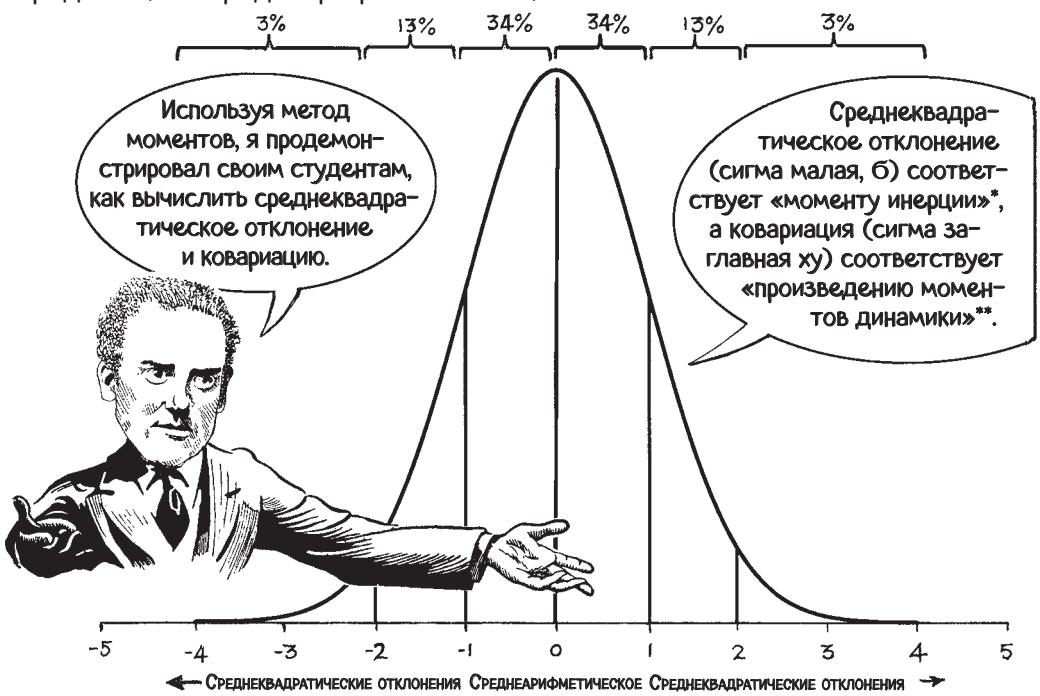
Преимущество размаха состоит в его простоте, однако это наименее надежный способ измерить изменчивость, так как он не использует всех имеющихся данных и подвержен влиянию выпадающих значений.

В данном примере речь идет о температурах по Цельсию одной недели ноября: 2, 6, 8, 12, 10, 12, 26. Размах равен 26–2=24 градусам.

Результат — 24 градуса — это необъективная и ненадежная численная оценка всего диапазона температур недели ноября. Нетипично высокая для такого сезона температура в 26 градусов является аномалией (или, возможно, одним из фактов глобального потепления).

Среднеквадратическое отклонение

Пирсон ввел понятие **среднеквадратического отклонения** в своей лекции в Грешем-колледже 31 января 1893 года, обозначая его сперва как «среднеквадратическое расхождение» (Standard divergence). Джон Венн использовал термин «расхождение» за несколько лет до этого, когда говорил об отклонении. Среднеквадратическое отклонение есть мера изменчивости. Оно показывает, как, широко или узко, разбросаны значения в наборе данных, а также показывает, как сильно отдельные значения отличаются от среднего (т. е. среднеарифметического).



Ковариация измеряет, как сильно две случайные величины соотносятся друг с другом. Если две величины движутся в одном направлении, то ковариация считается положительной. Если две величины движутся в разных направлениях, ковариация считается отрицательной. Если у двух величин относительно друг друга нет определенного направления, тогда ковариация равняется нулю.

---

* Момент инерции — это важный элемент механики. Это геометрическое свойство стержня, и он измеряет способность стержня сопротивляться сгибанию и деформации. — *Прим. науч. ред.*

** Момент динамики связан с действием силы, направленной на движение объектов. — *Прим. науч. ред.*

Используя среднеквадратическое отклонение, Пирсон добился того, что смог измерить *все* точки изменчивости на кривой распределения — вместо двух или трех точек, которые сумел измерить Гальтон с помощью квартильного размаха.

$$\text{Среднеквадратическое отклонение} = \sqrt{\frac{(\text{сумма исходных значений из набора данных} - \text{среднеарифметическое из этих данных наблюдений})^2}{\text{число наблюдений}}}$$

или:

$$S = \sqrt{\frac{\Sigma(X - \bar{X})^2}{N}}$$

Таким образом, среднеквадратическое отклонение равняется квадратному корню из средних отклонений, возведенных в квадрат.

1) Вычитаем среднеарифметическое из исходных данных (X), что дает нам значения «отклонения» (обозначается строчной x).
2) Чтобы исключить отрицательные значения, возводим в квадрат полученные значения.
3) Суммируем значения отклонения, возведенные в квадрат, и делим на число наблюдений, чтобы вычислить среднеквадратическое отклонение.

Формула:

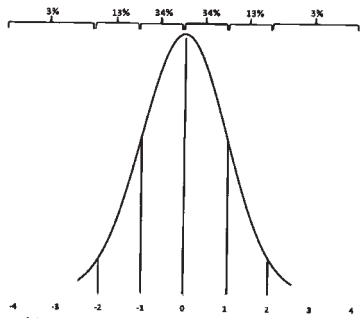
Исходные данные	– Средне-арифмети-ческое	= Значение отклонения	Значение отклонения в квадрате
X	$\bar{x}$	**x**	$\mathbf{x^2}$
12	8	4	16
10	8	2	4
6	8	-2	4
8	8	0	0
4	8	-4	16
$\frac{40}{5} = 8$			40

Формула отклонения:

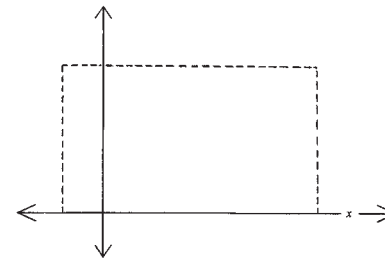
$$S = \sqrt{\frac{\sum x^2}{n}} = \sqrt{\frac{40}{5}} = \sqrt{8} = 2{,}82$$

Это значит, что средняя величина отклонения в указанном наборе данных на 2,82 единицы отстоит (находится в радиусе) от среднеарифметического значения 8 и что, следовательно, изменчивость в нашей выборке не так велика.

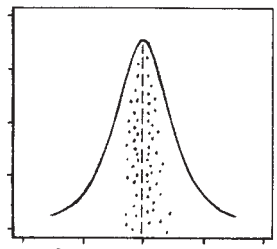
Большое среднеквадратическое отклонение (относительно значения среднеарифметического) показывает, что частотное распределение имеет большой разброс по значениям относительно среднеарифметического, в то время как малое среднеквадратическое отклонение показывает, что бóльшая часть значений сгруппирована рядом со среднеарифметическим, с небольшими отличиями в наблюдениях друг от друга. Несмотря на то, что среднеквадратическое отклонение показывает, насколько величины отличаются от среднеарифметического, по нему нельзя сказать о том, насколько величины целой группы значений отличаются друг от друга в более частной группе значений.



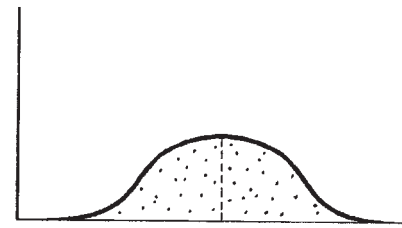
Малое среднеквадратическое отклонение



Большое среднеквадратическое отклонение



Среднеарифметическое



Среднеарифметическое

Если среднеквадратическое отклонение является практическим инструментом измерения изменчивости, то дисперсия используется в теоретических работах, в особенности в *дисперсионном анализе* (см. с. 168–170).

ДИСПЕРСИЯ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

ДИСПЕРСИЯ СЛУЧАЙНОЙ ВЕЛИЧИНЫ ДИСПЕРСИЯ СЛУЧАЙНОЙ ВЕЛИЧИНЫ ДИСПЕРСИЯ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

Дисперсия также измеряет изменчивость, однако она используется для случайных величин и обозначает степень разброса значений относительно ожидаемых (а не фактических) значений*.

Используя пример, рассмотренный для среднеквадратического отклонения:

Дисперсия случайной величины =

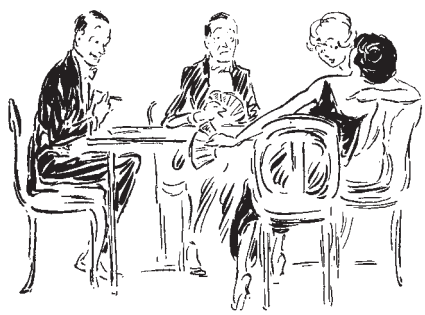
$$\frac{(\text{сумма исходных значений из набора данных} - \text{среднеарифметическое из наблюдений})^2}{\text{число наблюдений}}$$

или

$$S^2 = \frac{\Sigma(X - \bar{x})^2}{N}.$$

Формула отклонения для дисперсии случайной величины:

$$S^2 = \frac{\Sigma x^2}{N} = \frac{40}{5} = 8$$

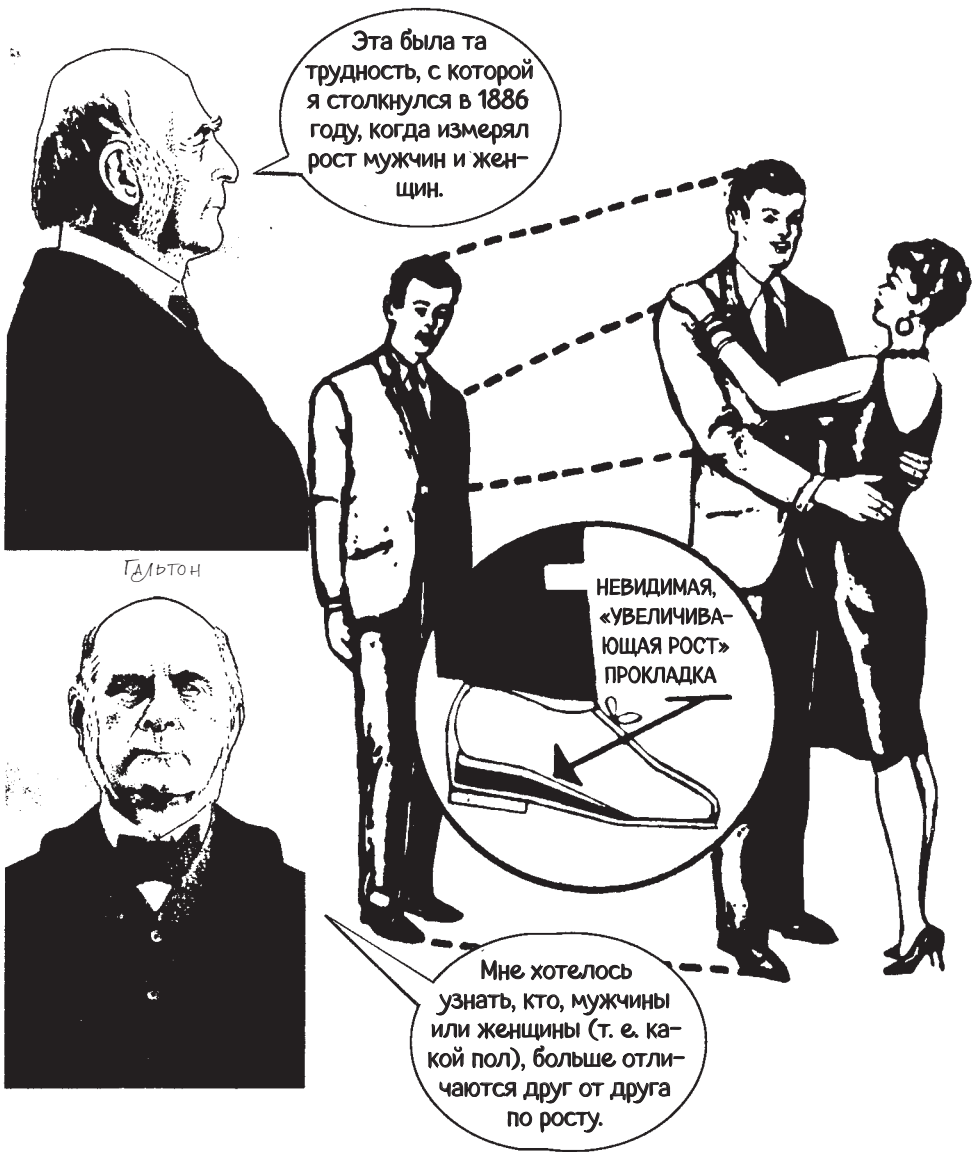


* Ожидаемые значения (expected values), или в узком смысле математическое ожидание — это среднее значение случайной величины, которое появляется (ожидается) в процессе испытаний, повторенном много раз, и при наличии идентичных шансов осуществления соответствующих исходов (реализации случайной величины).

ДИСПЕРСИЯ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

Так как среднеквадратическое отклонение не отражает диапазона изменчивости (range of variation) внутри этой группы, то как Пирсон смог определить, насколько изменяются значения внутри группы, и как он смог провести сравнения с другими группами, в которых получаются иные значения среднеарифметического?

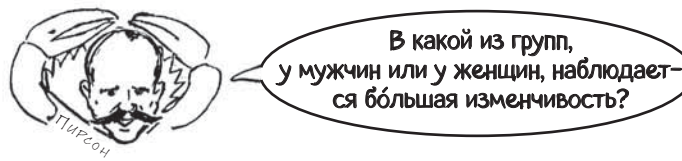
Для этого требуется другой статистический метод.



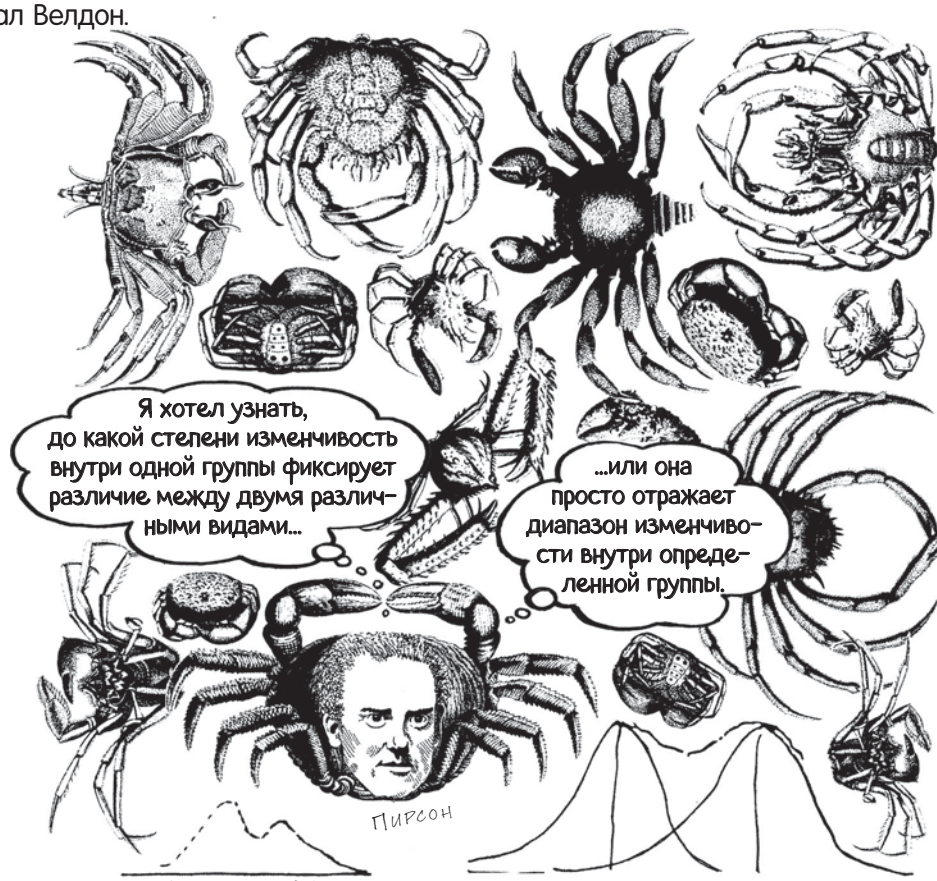
Гальтон справился с этой трудностью, соединив среднее телосложение женщин с эквивалентным ему средним телосложением мужчин, а затем сравнив выравненные (сопоставимые) отклонения у мужчин и у женщин. Выравнять или «преобразовать» средний рост женщин в средний рост мужчин ему удалось, умножив женский рост на константу, равную 1,08.

Вариационный коэффициент

Пирсон думал, что наилучшим способом сравнивать отклонения в росте женщин и мужчин был следующий: нужно варьировать отклонения в одинаковой пропорции. Использование одного только метода среднеквадратического отклонения, который позволял измерять сантиметры или дюймы, скорее всего показало бы, что мужчины в среднем выше, так как имеют большее среднеарифметическое значение роста. Однако метод не отвечает на такой вопрос:

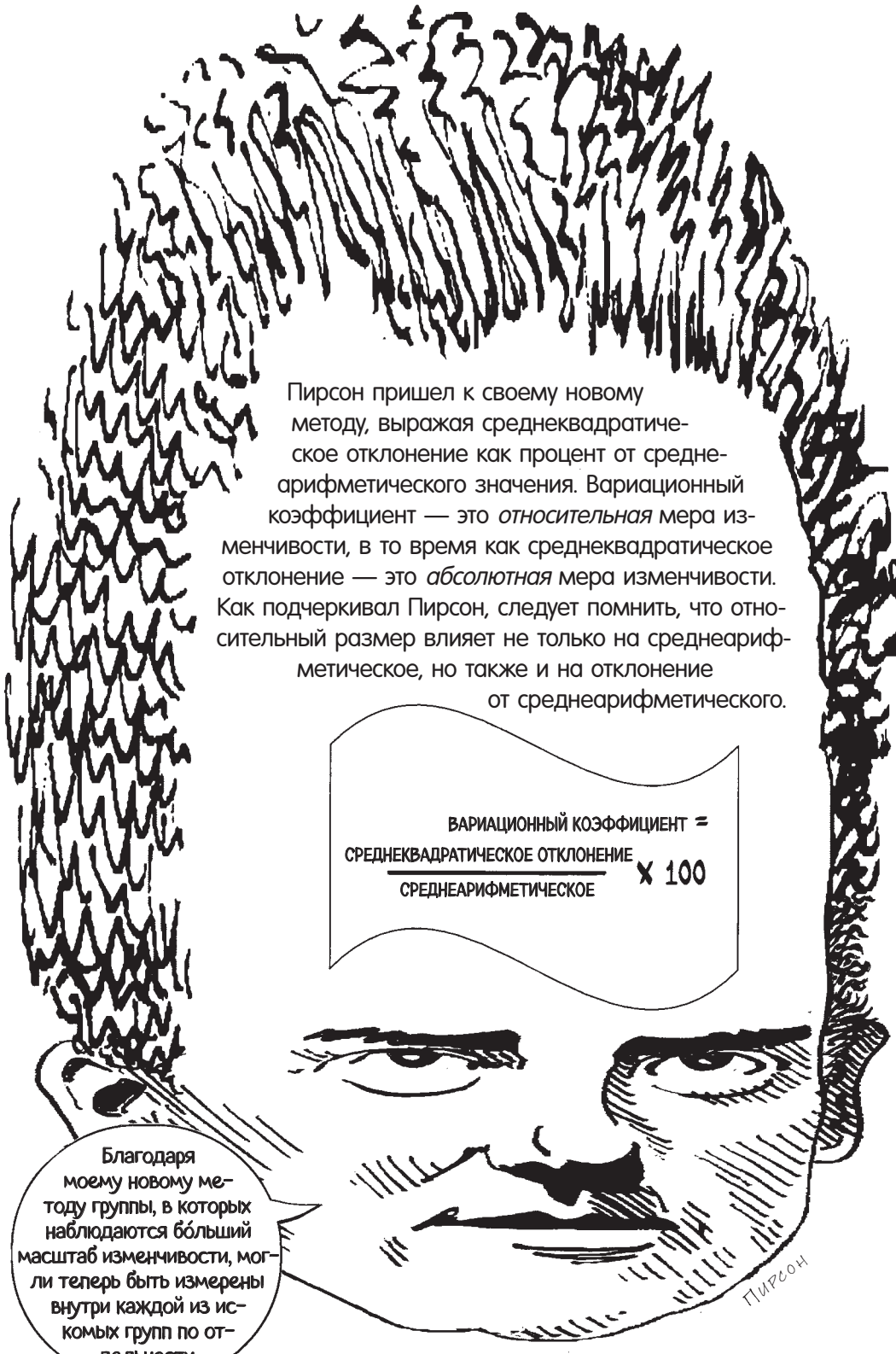


Пирсон придумал **вариационный коэффициент** для ответа на заданный вопрос. Это было важно для Пирсона еще и потому, что он пытался определить, насколько изменчивым было поведение креветок и крабов, с которыми работал Велдон.



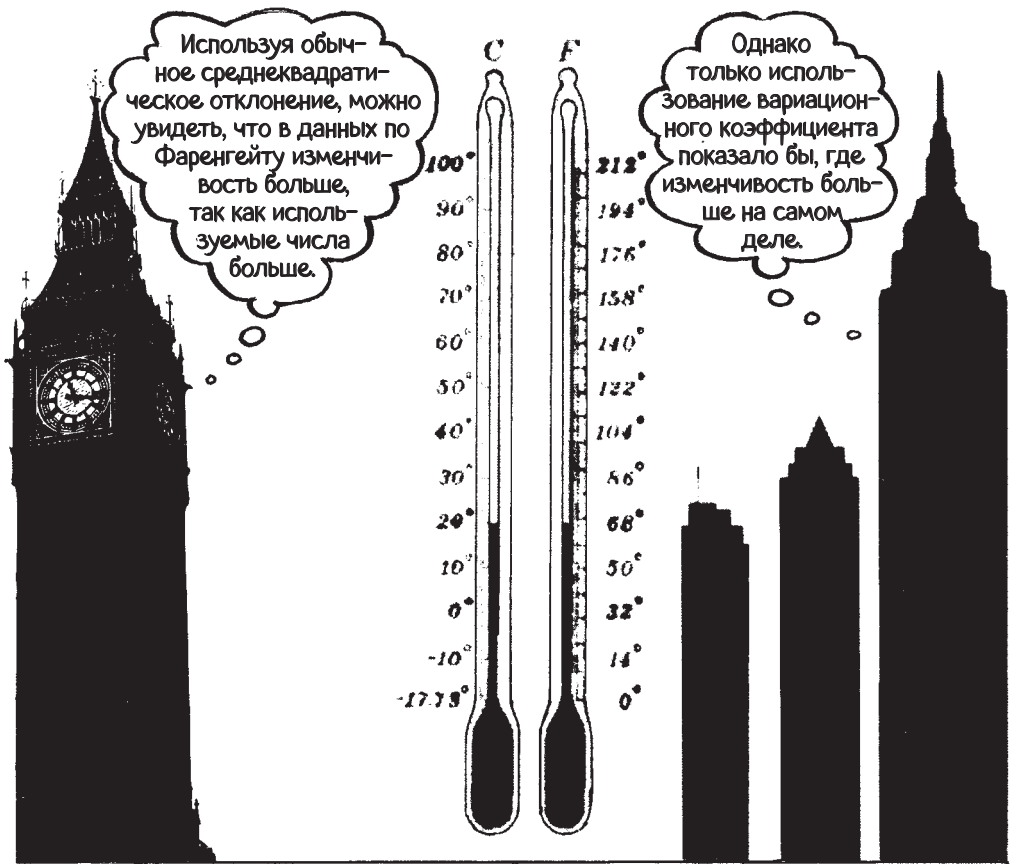
КРИВАЯ ВЕЛДОНА С ДВУМЯ МАКСИМУМАМИ

РАЗБИЕНИЕ ДВУХ КРИВЫХ НОРМАЛЬНОГО РАСПРЕДЕЛЕНИЯ, ПРОИЗВЕДЕННОЕ ВЕЛДОНОМ


Пирсон пришел к своему новому методу, выражая среднеквадратическое отклонение как процент от среднеарифметического значения. Вариационный коэффициент — это *относительная* мера изменчивости, в то время как среднеквадратическое отклонение — это *абсолютная* мера изменчивости. Как подчеркивал Пирсон, следует помнить, что относительный размер влияет не только на среднеарифметическое, но также и на отклонение от среднеарифметического.
ВАРИАЦИОННЫЙ КОЭФФИЦИЕНТ = СРЕДНЕКВАДРАТИЧЕСКОЕ ОТКЛОНЕНИЕ / СРЕДНЕАРИФМЕТИЧЕСКОЕ x 100
Благодаря моему новому методу группы, в которых наблюдаются бóльший масштаб изменчивости, могли теперь быть измерены внутри каждой из искомых групп по отдельности.
ПИРСОН

Сравнивая изменчивость величин

Вариационный коэффициент не имеет единиц измерения, поэтому его можно использовать при сравнении изменчивости разных величин с разными единицами измерения. Следовательно, сравнения могут быть проведены для градусов Цельсия в Лондоне и Фаренгейта в Нью-Йорке за одну неделю, для того чтобы определить, где изменчивость температур больше.



Лондон	Нью-Йорк
Градусы Цельсия	**Градусы Фаренгейта**
15	40 Понедельник
19	60 Вторник
20	70 Среда
13	55 Четверг
24	75 Пятница
18	65 Суббота
21	70 Воскресенье

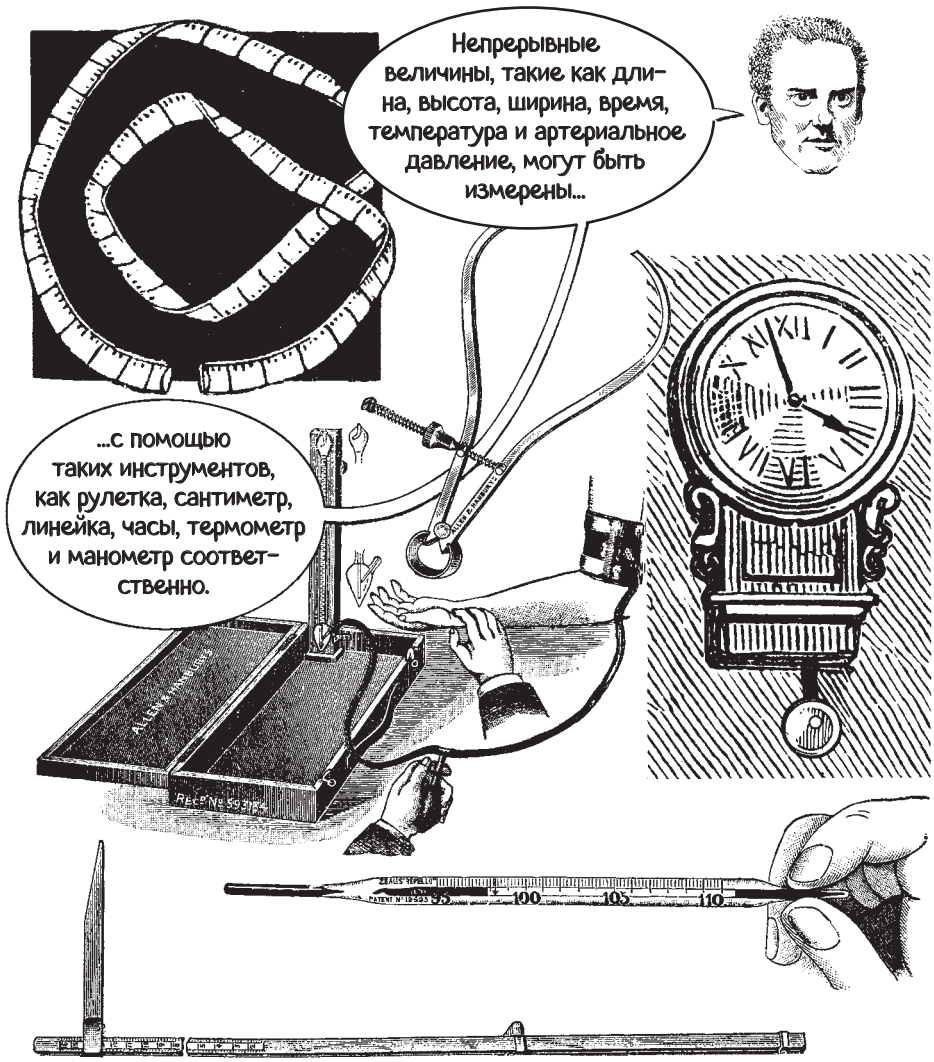
Практическое применение

Этот метод продолжает широко использоваться в производстве, маркетинге и экономической науке. Производители шерсти применяют вариационный коэффициент для вычисления изменчивости в распределении диаметра волокна и неоднородности пряжи.

Например, эта информация позволяет производителям создавать различную по качеству шерсть, в зависимости от требований рынка.

Шкалы измерения Пирсона

Различение шкал измерения было очень важно в развитии как методов корреляции Пирсона, так и других статистических тестов. Когда Гальтон, Велдон и Пирсон первыми начали анализировать статистические данные, практически все они были непрерывными. К 1899 году Пирсон начал работу над статистическими коэффициентами для измерения соотношений между «прерывными», или дискретными, величинами.

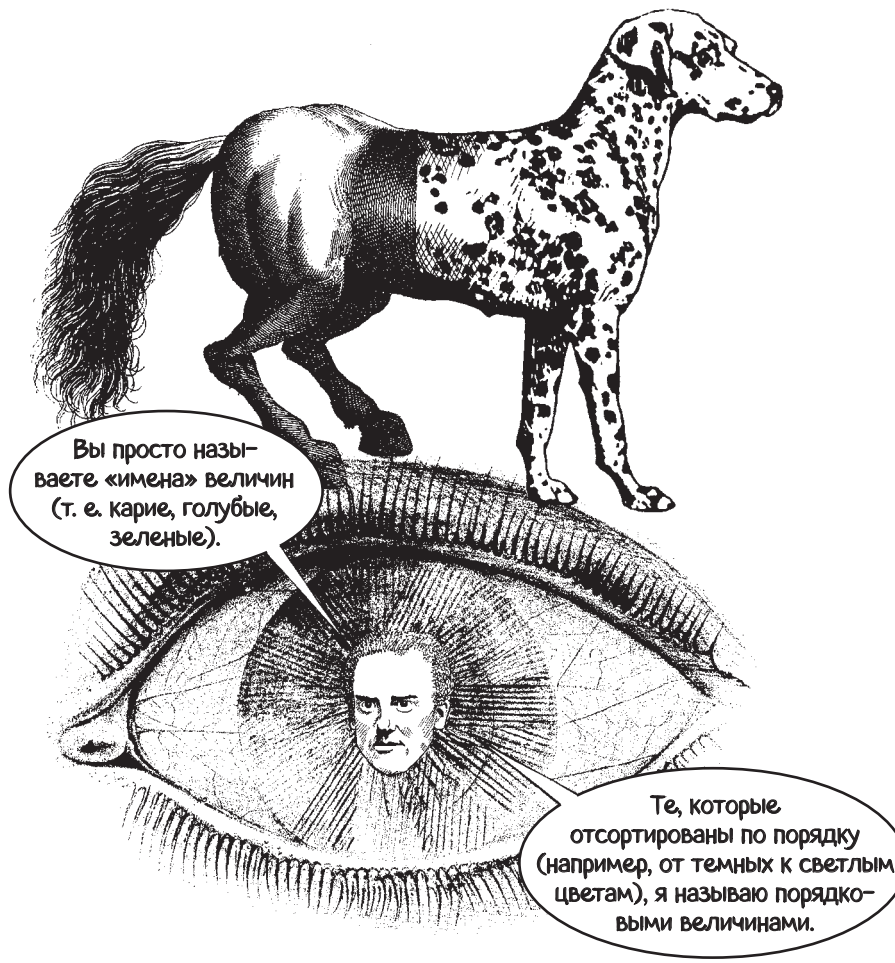


Эти величины выражены в таких единицах измерения, которые могут быть представлены в конечных единицах, таких как дюймы, сантиметры, секунды, минуты и градусы.

Номинальные и порядковые величины

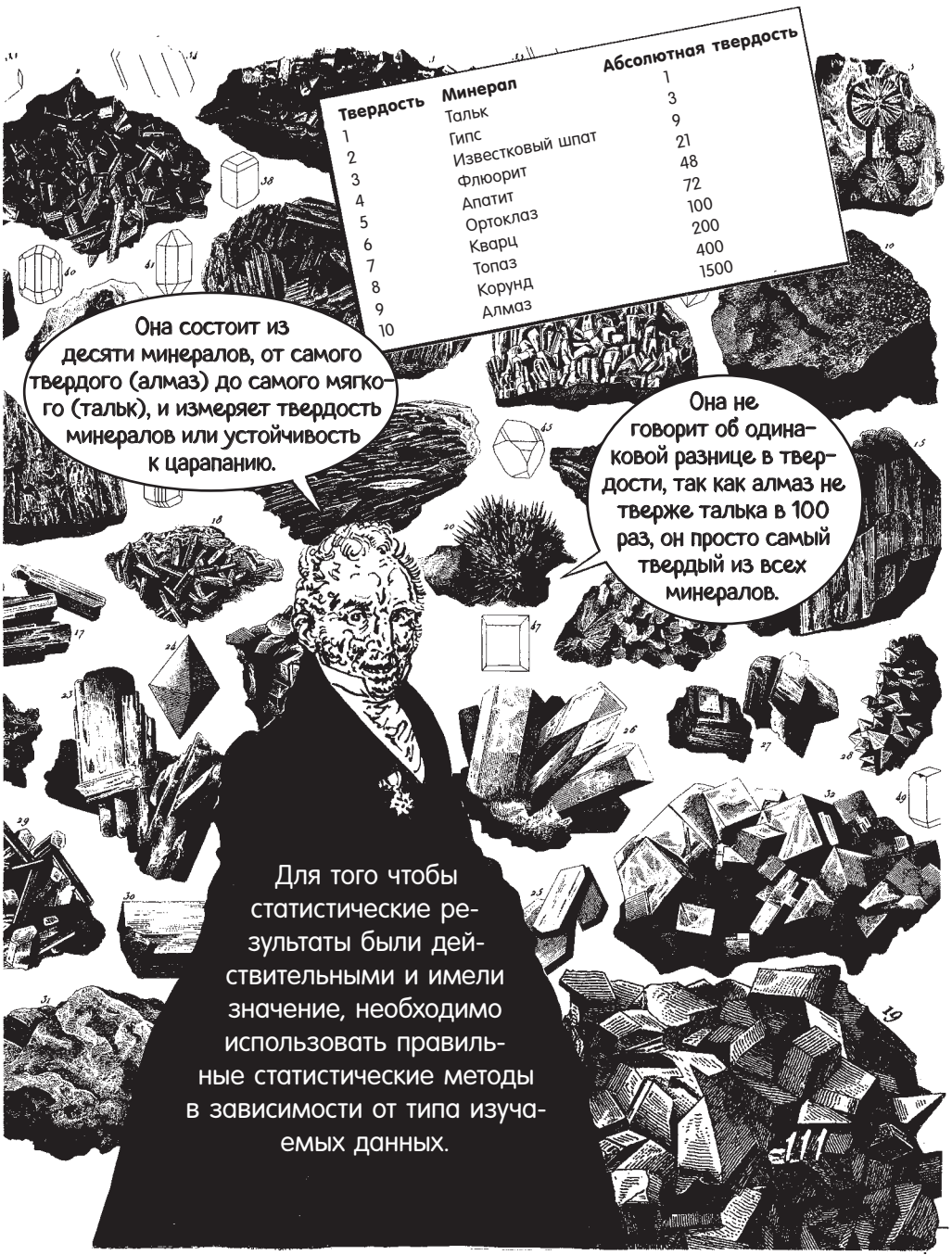
Пирсон впервые столкнулся с величинами, с которыми нельзя было работать как с непрерывными, когда он начал изучать наследование цвета глаз у людей и окрас шерсти у лошадей и собак. В этих исследованиях единственная доступная форма классификации величины — эта та, которая включала «подсчет», а не «измерение»: цвет глаз не может быть измерен тем же способом, что и телосложение, вес или время.

Пирсон назвал такие величины, как цвет глаз, **номинальными**.



Номинальные величины включают почти все демографические величины, такие как вероисповедание, политические убеждения и социально-экономический статус.

Порядковые величины вначале сортируются, а затем именуются. Шкала Мооса (Mohs) [минералогическая шкала твердости], придуманная немецким минерологом Фридрихом Моосом в 1822 [на самом деле в 1811] году, является примером порядковой шкалы.



Соотношение и интервал

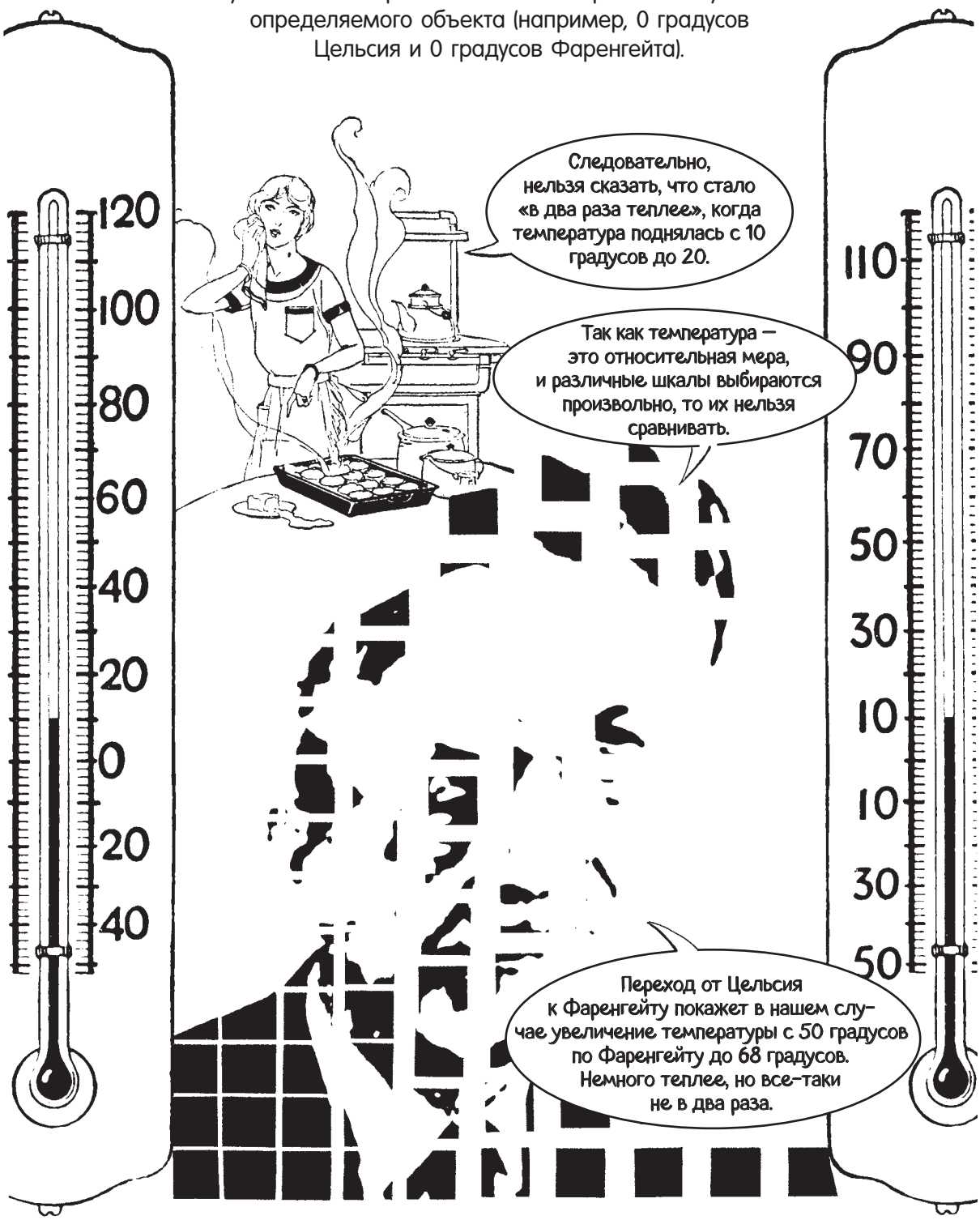
Американский психолог **Стенли Смит Стивенс** (Stevens) (1906–1973) произвел более глубокое подразделение внутри «непрерывных величин» в 1947 году, когда ввел понятие *интервальной* шкалы и шкалы *отношений* (ratio scales) (большая часть непрерывных величин Пирсона относилась к шкале *отношений*). Стивенс предложил следующее.

1. Шкала отношений

Она отличается от интервальной шкалы (см. следующую страницу) в двух пунктах: а) абсолютный нуль обозначает отсутствие свойства измеряемой величины (т. е. высоты, веса и артериального давления) и б) шкала отношений аддитивна.

2. Интервальная шкала

Нулевая точка произвольна и не отражает отсутствие определяемого объекта (например, 0 градусов Цельсия и 0 градусов Фаренгейта).



Корреляция

Корреляция является одним из наиболее широко используемых статистических методов, обозначающих степень, до которой две величины идут вместе (например, высота и вес). Наиболее частый тип корреляции измеряет линейные отношения между двумя величинами и обозначает, как близко они идут друг с другом относительно прямой линии.

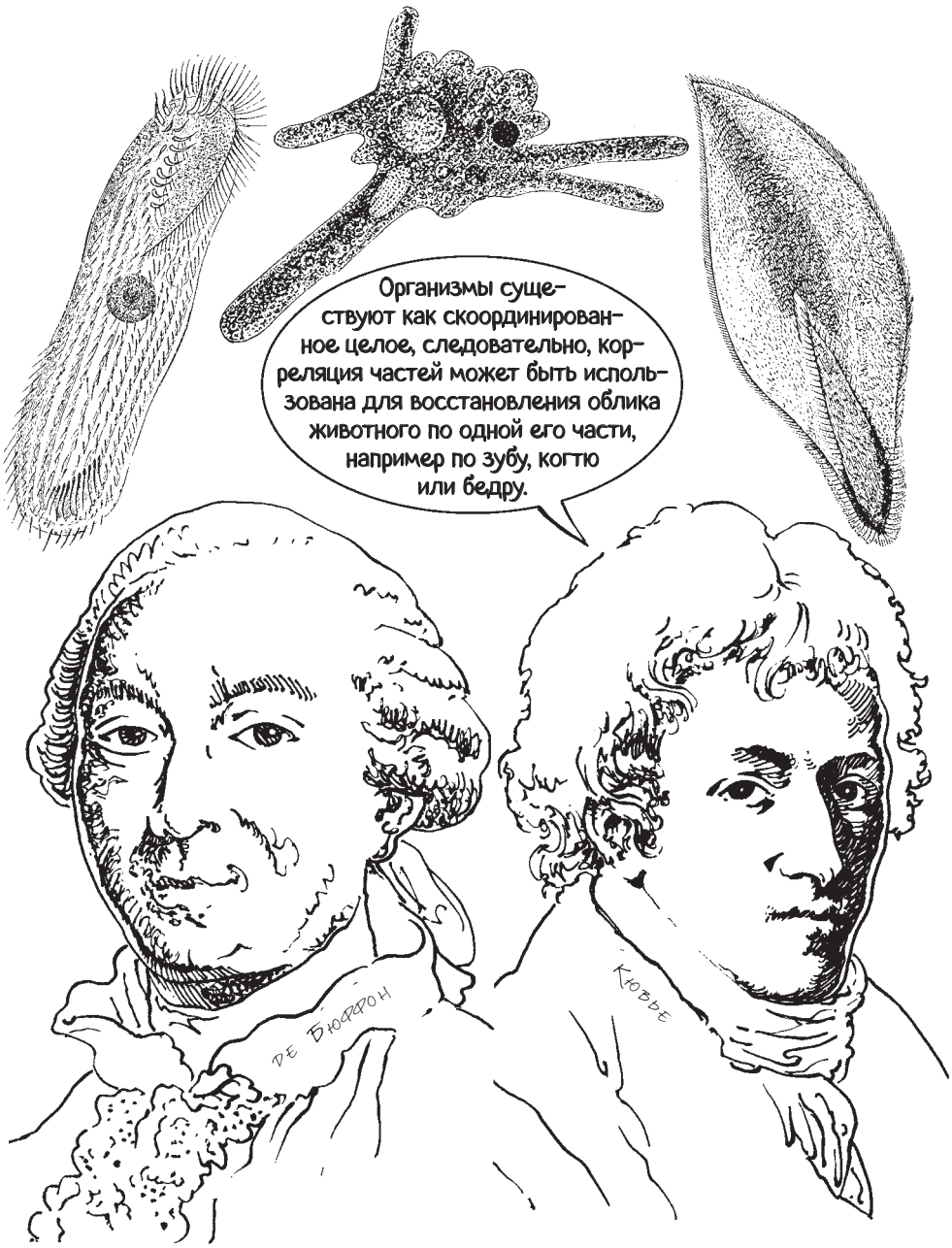
Однако не каждую пару характеристик или величин можно соотнести, используя статистическую корреляцию. Различные способы корреляции используются в биологических, медицинских, поведенческих, социальных и естественных науках, так же как и в производстве, торговле, экономической науке и образовании.

Различные типы корреляции используются для различных типов величин, в зависимости от шкалы измерения.

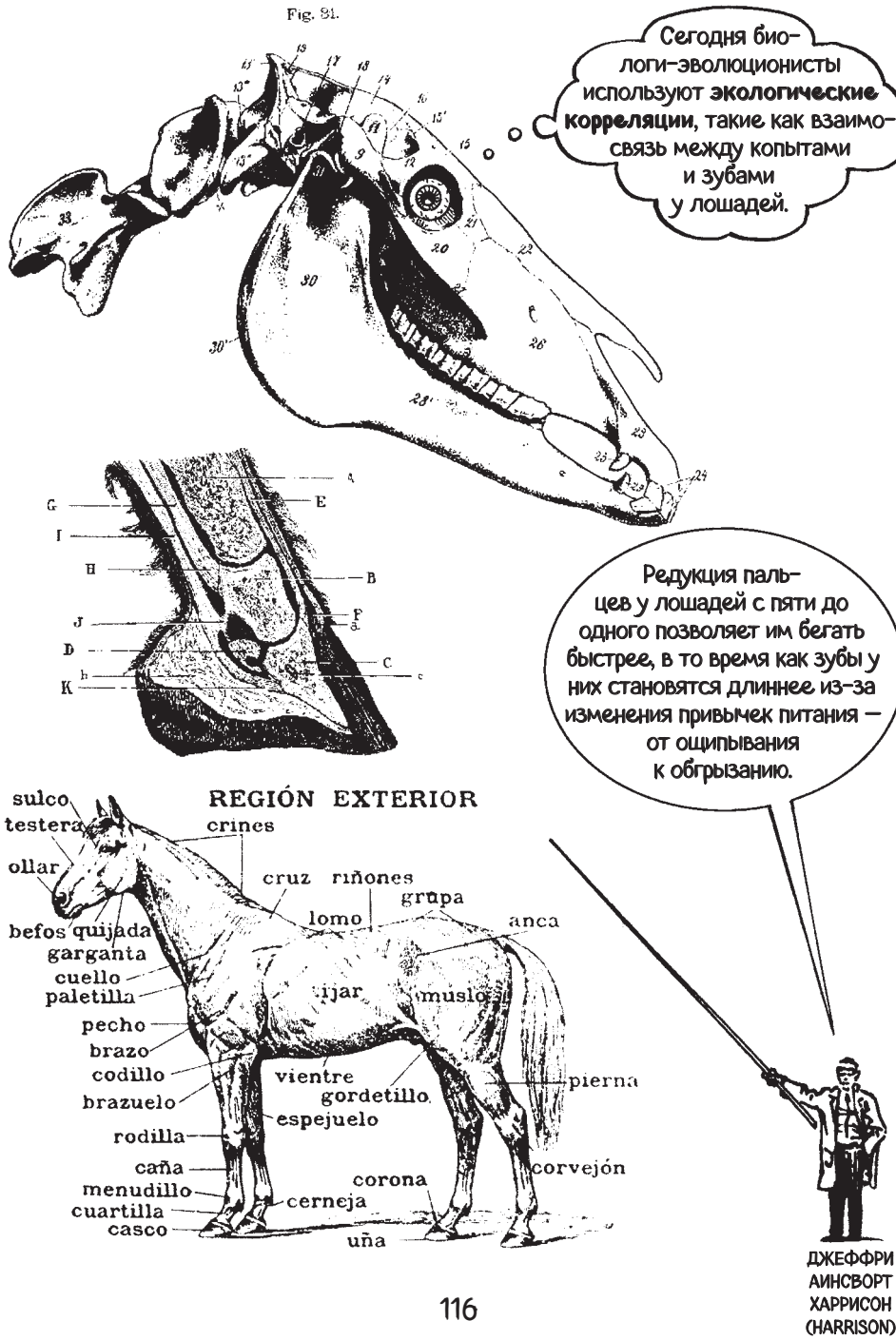
Пирсон изобрел методы для всех типов переменных.

Раннее использование корреляции

Термин «корреляция» был уже в ходу целый век, прежде чем был найден способ измерить ее. Первым, кто использовал этот термин, был биолог **граф де Бюффон** (Buffon) (1707–1788), а затем понятие о корреляции было развито палеонтологом **Жоржем Леопольдом Кювье** (Cuvier) (1769–1832), который писал о «корреляции частей» в 1801 году.



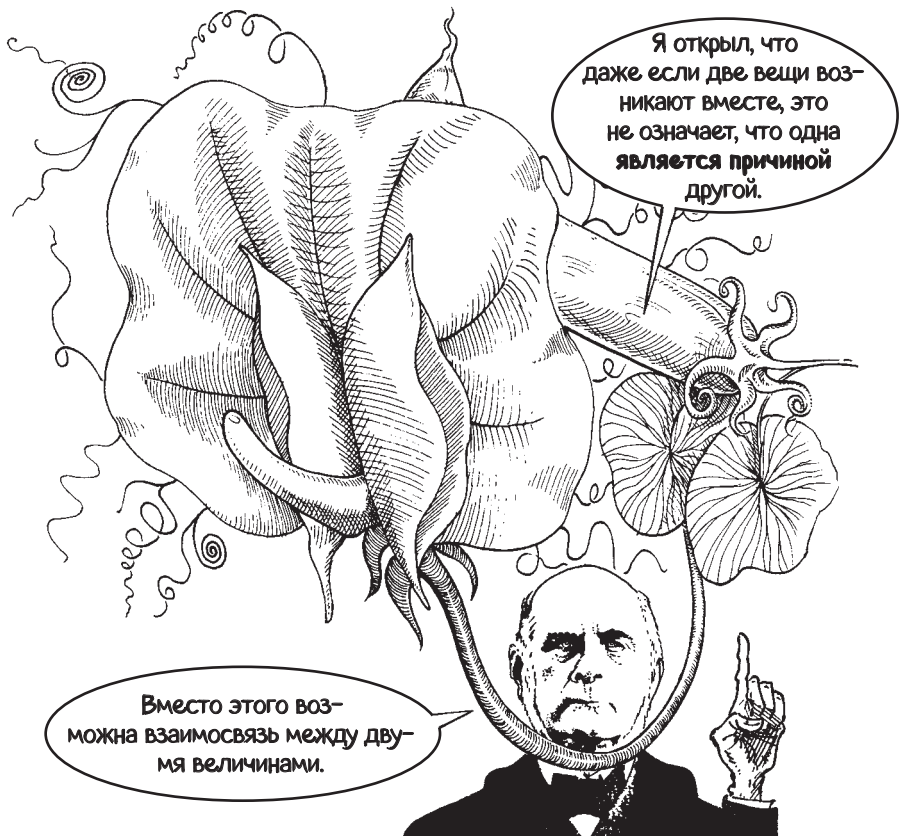
Чарлз Дарвин считал идею Кювье о корреляции частей важной и полезной и говорил о *функциональных корреляциях,* когда, например, размер одного органа является функцией другого органа. Дарвин также говорил о *корреляции развития,* которая появляется на ранних стадиях роста и влияет на развитие организма.



Причинность и ложная корреляция

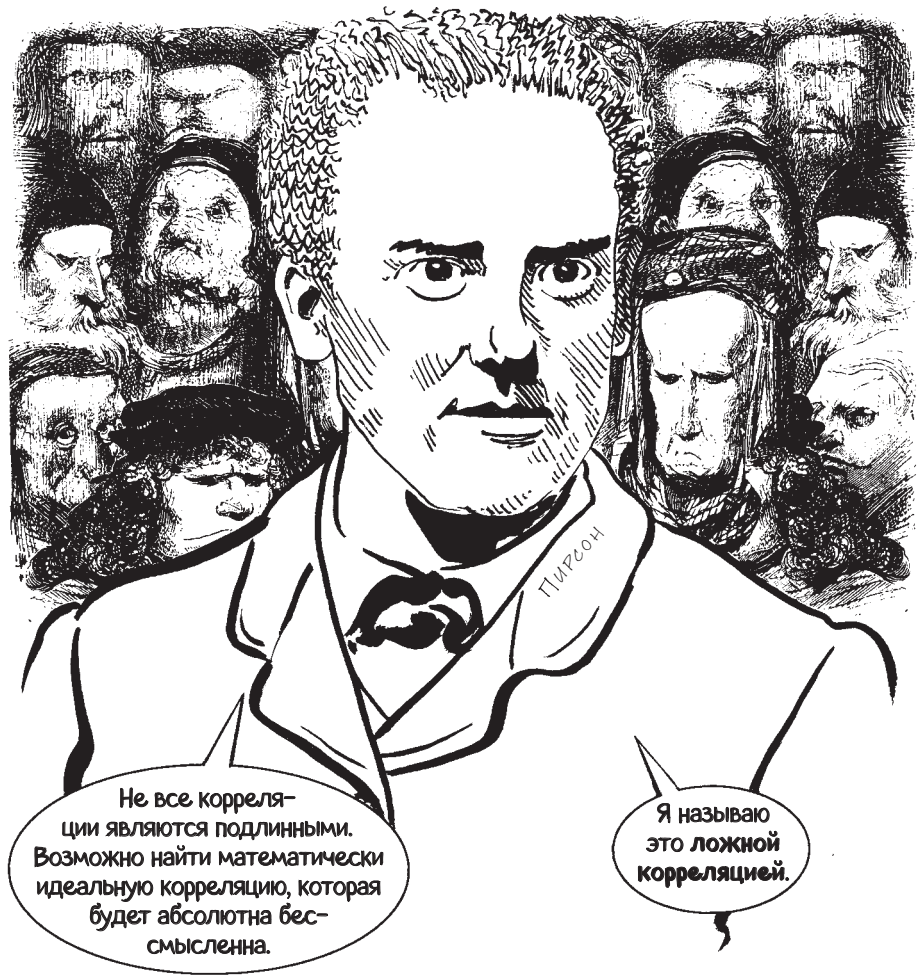
Фрэнсис Гальтон был первым человеком, который придумал способ измерения корреляции. Он создал график для поиска взаимосвязи между мамой и дочкой душистого горошка.

До того как Гальтон придумал идею корреляции, **причинность** была основным способом, которым объяснялись два связанных между собой события, в особенности в естественных науках.



До встречи с Гальтоном Пирсон был убежден, что формально математика может быть применима только к явлениям природы; эти последние определялись бы тогда причинностью. Однако идеи Гальтона о корреляции заменили Пирсону причинность, в особенности это касалось биологических наук. Он стал противником причинно-следственных связей, считал, что Вселенная управляется не законами причинно-следственной связи в своей узкой форме, а больше изменчивостью, которой отводилась крупная роль в объяснении явлений.

Пирсон предупреждал своих студентов, что корреляцию не следует понимать как признак причинности. Хотя он и осознавал, что «для тех, кто настаивает на сведе́нии всех корреляций к действию лишь причин и следствий, тот факт, что корреляция может быть установлена и между двумя не имеющими связи явлениями, такое положение дел может оказаться шоком». Более того, направление причинности неизвестно: X вызывает Y или Y вызывает X?

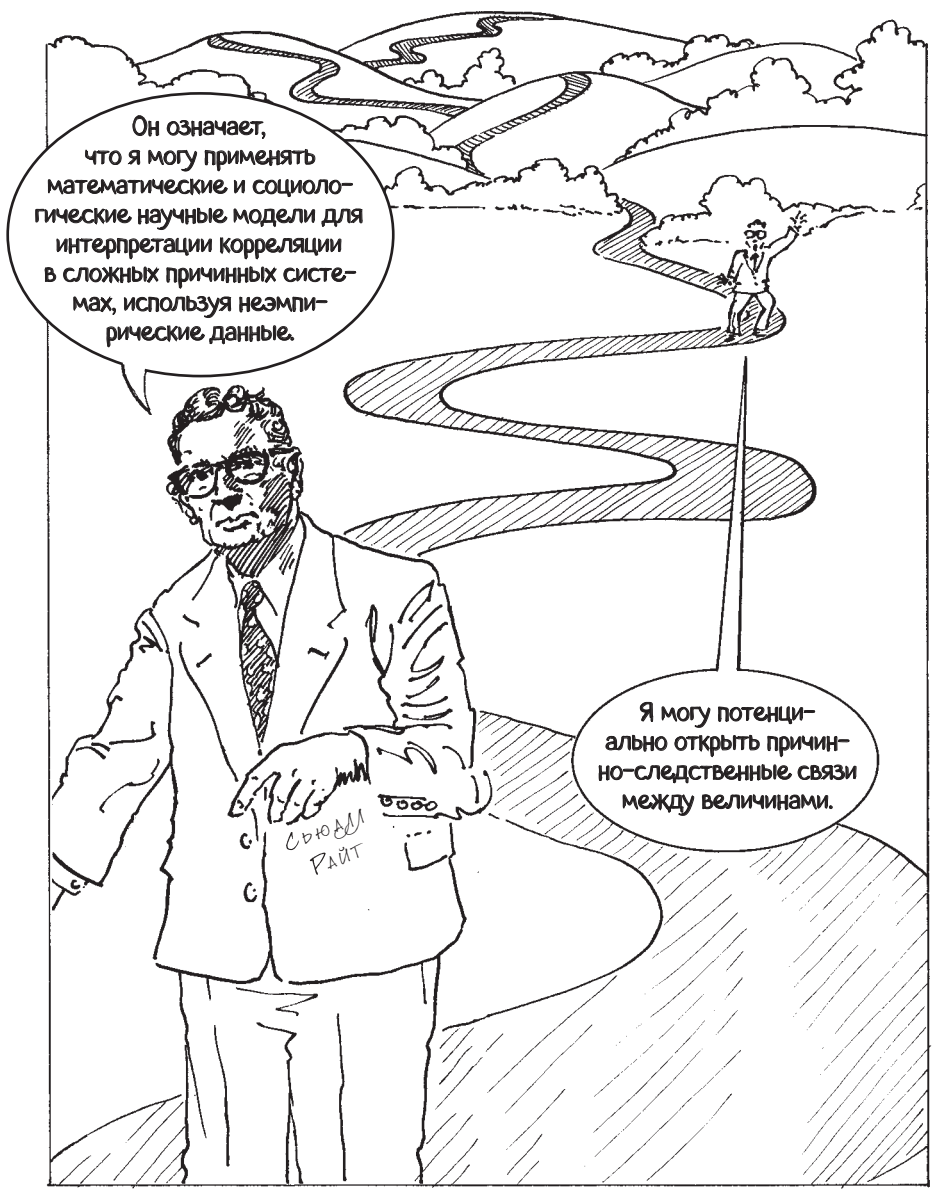


Следовательно, математически идеальная корреляция не означает причинности: она просто значит, что две величины очень сильно коррелируют друг с другом. Такой результат может получаться и в случае ложной иллюзорной (или кажущейся) корреляции из-за влияния третьей величины, называемой скрытой величиной. Если квалификация студентов университетов сильно коррелирует с их последующими доходами (чем лучше знания, тем выше зарплата), эта корреляция может возникать вследствие третьей (скрытой, или неявной) величины, например, способности усердно трудиться.

Пат-анализ и причинность

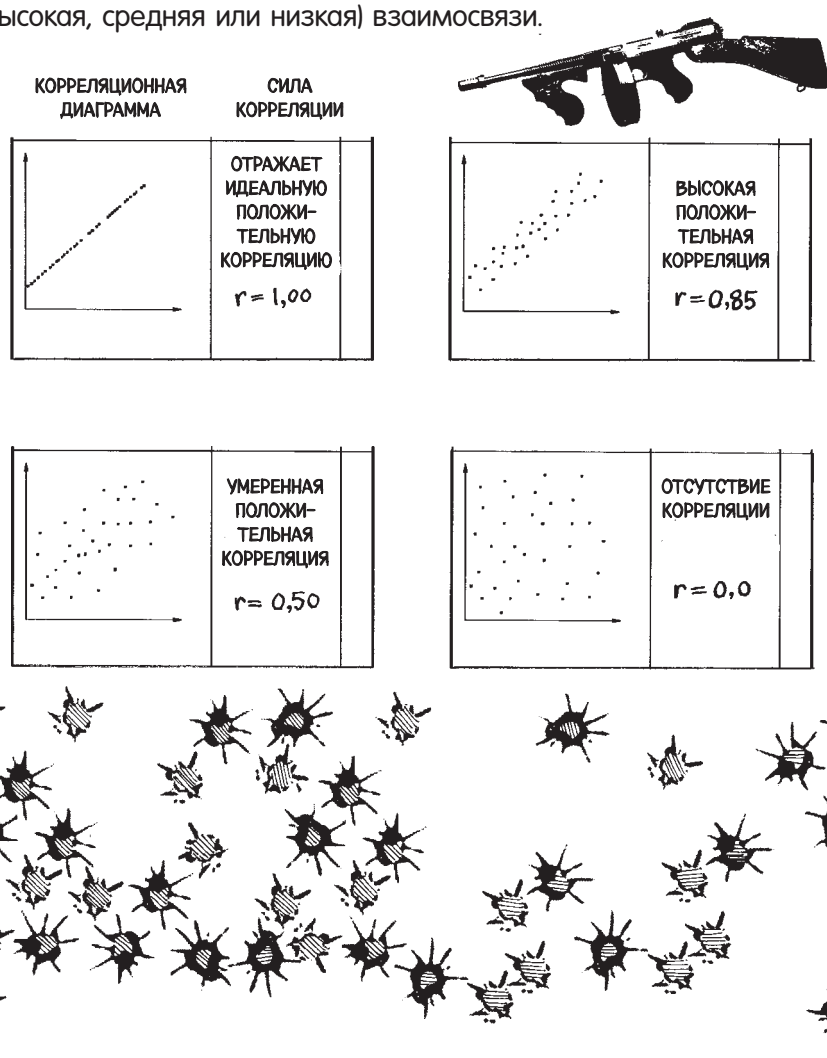
Биолог-эволюционист Сьюалл Райт (Wright) расширил идеи Пирсона о корреляции в области причины и следствия, изобразив на рисунке логические и методологические взаимосвязи между корреляцией и причинностью.

Используя пирсоновскую множественную регрессию (см. с. 134–138), в 1918 году Райт придумал статистическую методологию, которую он назвал *пат-анализом*.



Корреляционные диаграммы, или диаграммы рассеяния

Корреляция часто изображается графически в виде так называемых корреляционных диаграмм для того, чтобы увидеть ее форму. Если две величины дают узкий эллипс, который напоминает прямую линию, это будет означать высокую корреляцию. Обычный эллипс говорит о средней корреляции, а круг обозначает отсутствие корреляции. С помощью такого способа измеряется сила (высокая, средняя или низкая) взаимосвязи.



Однако корреляцию невозможно перевести в проценты. Следовательно, умеренная корреляция 0,55 или высокая 0,80 не соответствует 55% или 80%, как ошибочно полагают некоторые люди.

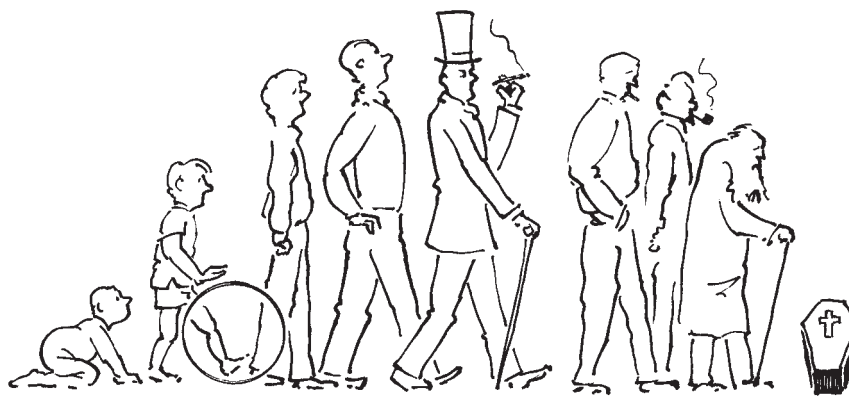
Велдон и отрицательная корреляция

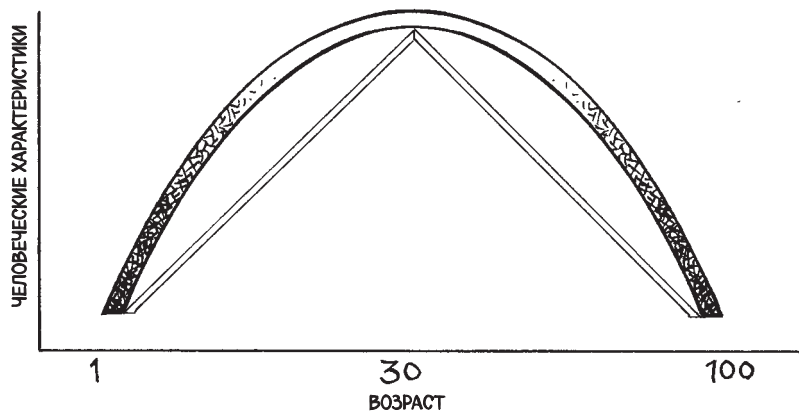
Числовой индекс, который дает корреляция, также показывает *направление* взаимосвязи. Две величины могут или расти, или убывать одновременно (например, рост и масса тела здоровых детей увеличиваются), или же одна величина растет, а вторая убывает (например, чем быстрее едешь на машине, тем скорее прибываешь в пункт назначения, т. е. скорость растет, время убывает). Первый процесс дает положительную, или прямую, корреляцию, а второй дает отрицательную, или обратную, корреляцию.

Взаимосвязи переменных, представленные разными кривыми

Несмотря на то, что численный индекс предоставляет информацию о степени линейной взаимосвязи, корреляционная диаграмма, или диаграмма рассеяния — это полезный инструмент, потому что он способен показать взаимосвязь переменных посредством кривых. В 1905 году Пирсон ввел в оборот корреляционное соотношение (correlation ratio) для измерения именно таких связей.

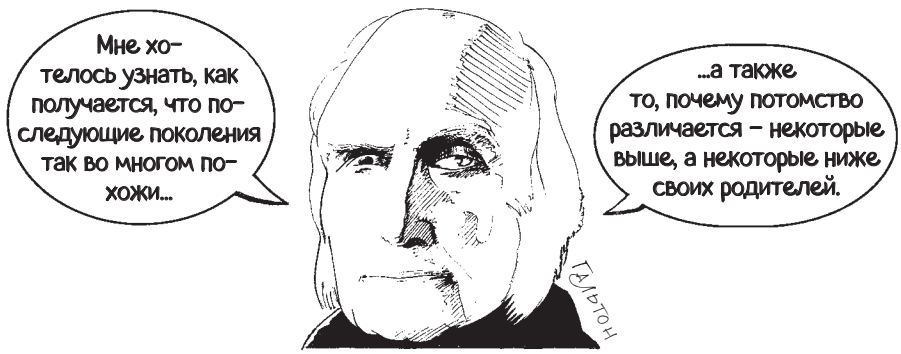
Возраст, соотнесенный с кривой роста в течение жизни, представляет собой довольно сложную кривую на графике, хотя кривая роста в детстве линейна. Дети продолжают расти до юности: они становятся выше, у них появляется волосяной покров, и сами они делаются более проворными, ловкими и гибкими. Однако продолжительность жизни представляет собой уже кривую, а не прямую, так как некоторые из этих характеристик сокращается с возрастом: рост уменьшается, мужчины имеют склонность к облысению, в целом люди становятся менее ловкими и гибкими в процессе старения.





Гальтон и биологическая регрессия

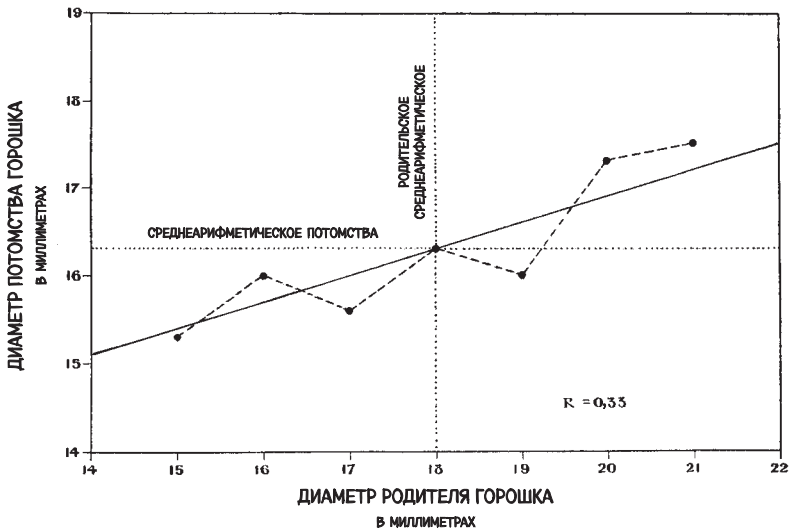
До своих работ по корреляции Гальтон занимался *регрессией*.



В 1875 году Гальтон измерял диаметр и вес тысяч пар матерей и дочерей душистого горошка и обнаружил, что популяция потомства возвращается к родителям и следует закону нормального распределения. Если размер матери-горошка увеличивался, то размер дочери-горошка также будет увеличиваться, но потомство не будет таким большим или таким маленьким, как мать-горошек. Следовательно, оно «регрессирует», оно возвращается назад к размеру «горошка-прародителя».

НАСЛЕДСТВЕННОСТЬ РАЗМЕРА СЕМЯН ДУШИСТОГО ГОРОШКА

Лекция Гальтона в Королевской ассоциации в 1877 году.



ГАЛЬТОНОВА ЛИНИЯ РЕГРЕССИИ ДУШИСТОГО ГОРОШКА

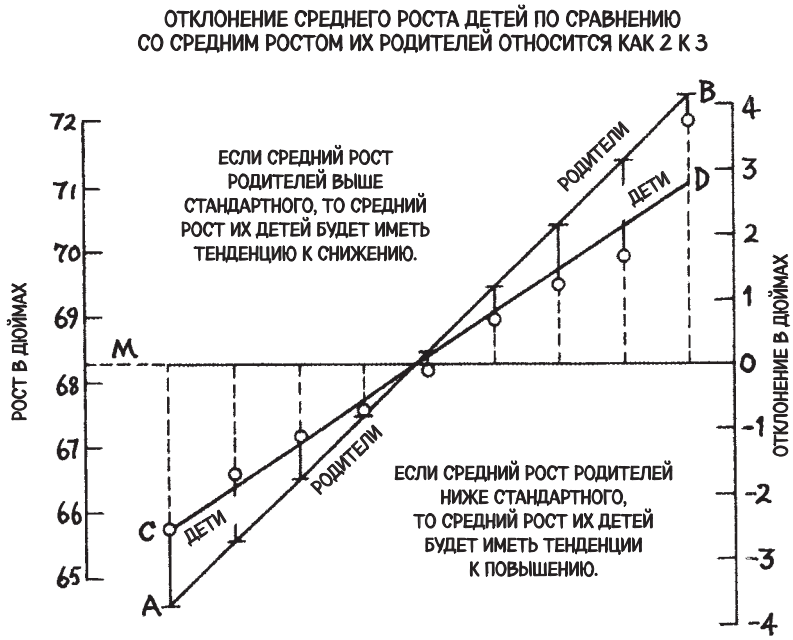
Регрессия к среднему значению

Она обозначает тенденцию какой-либо характеристики популяции сдвигаться от крайних значений ближе к средним.

Гальтон интересовался корреляцией роста у отцов и сыновей, так как ее было легко измерить и она оставалась устойчивой на протяжении взрослой жизни.

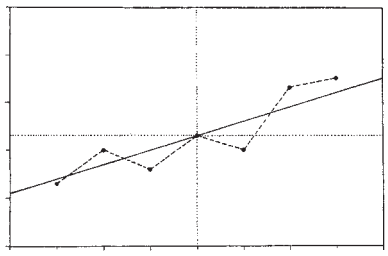
Я понял, что корреляция была в двух направлениях и порождала две кривые регрессии: одна от потомков к родителям и вторая от родителей к потомкам.

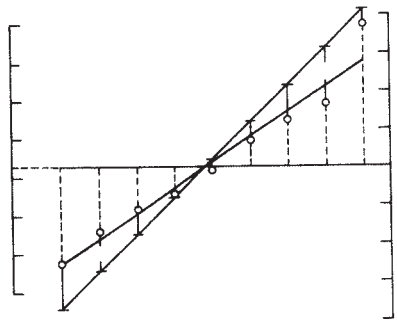
Однако результат Гальтона создал парадокс, который противоречил тому, что он понимал под односторонней регрессией. Гальтону пришлось объяснять, как рост потомков мог влиять на рост родителей.



Две кривые регрессии Гальтона

Пока Гальтон демонстрировал существование корреляции между отцами и сыновьями, две его кривые регрессии давали другую картину. Кривые на верхней части графика (см. с. 124) показывают, что если родители были выше среднего, то их дети будут ниже своих родителей: показатель среднего роста детей «регрессирует» к среднему значению. И наоборот, линии регрессии в нижней части графика показывают, что если родители были ниже среднего, то их дети будут выше своих родителей, показатель среднего роста будет также «регрессировать» к среднему значению.





Рост отцов и детей используется для иллюстрации отдельного случая регрессии к среднему значению.

Таблица А

Регрессия роста отца к росту сына

Отец = 185 см
Среднеарифметическое = 170 см
Сын = 179 см

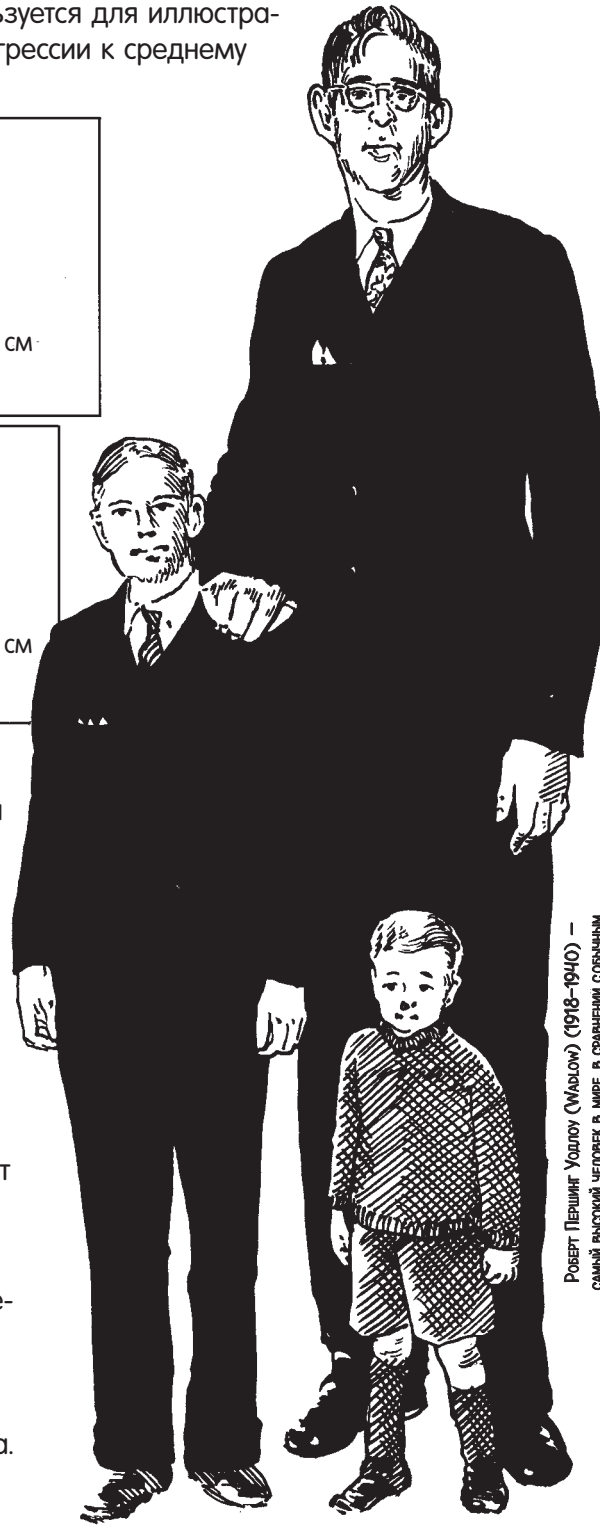
Таблица Б

Регрессия роста сына к росту отца

Сын = 188 см
Среднеарифметическое = 175 см
Отец = 170 см

В первой таблице средний рост взят из выборки 100 отцов и их сыновей и равняется 170 см, а рост отца равняется 185 см. Если происходит регрессия роста от отцов к сыновьям, то рост сына 179 см. Рост отца выше среднего, однако сын ниже отца, следовательно, значение регрессирует к среднему значению.

Во второй таблице рост сына равняется 188 см, регрессия роста отца ведет к отметке в 170 см. Здесь рост отца ниже среднего, однако его сын выше отца.



Роберт Першинг Уодлоу (Wadlow) (1918–1940) – самый высокий человек в мире, в сравнении собачным взрослым иребенком [его рост был 272 см].

Так как регрессия к среднему значению обозначает склонность характеристики популяции сдвигаться от крайних значений к средним, это укрепило мнение Гальтона о том, что распределение всегда будет нормальным. Он был убежден, что естественный отбор не мог создавать постоянные перемены в популяции, так как следующее поколение будет регрессировать к среднему значению вида.

Гальтон не учел тот факт, что последующее размножение и воспроизводство потомства после естественного отбора изменило форму распределения: кривая нормального распределения восстанавливается, но с другим средне-арифметическим значением (см. с. 93).

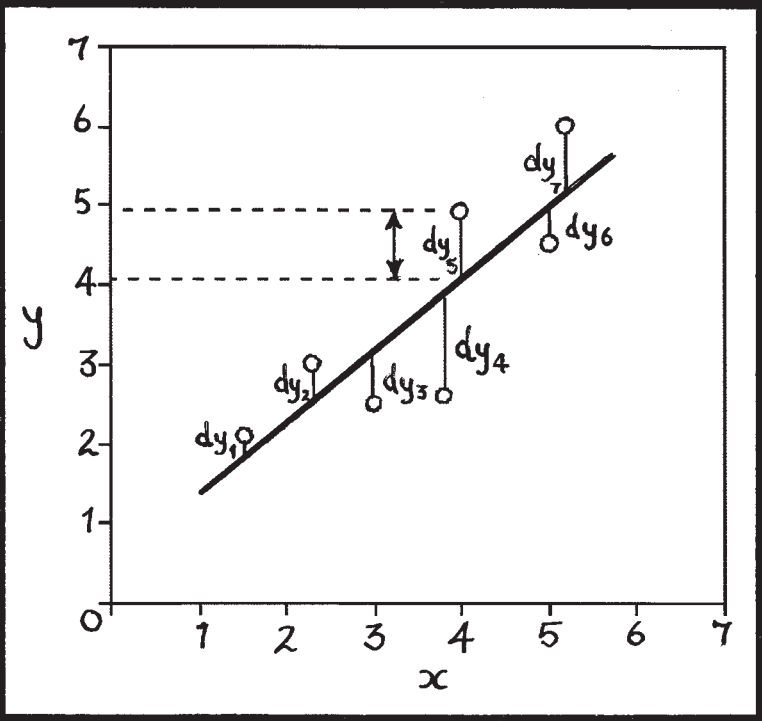
Однако регрессия не влияет на изменчивость (или дисперсию) популяции: изменчивость не уменьшается вследствие явления регрессии.

ЗАЛИТАЯ КРИВАЯ = ОРИГИНАЛЬНАЯ КРИВАЯ ДО ОТБОРА

ПУНКТИРНАЯ КРИВАЯ = НОВАЯ КРИВАЯ НОРМАЛЬНОГО РАСПРЕДЕЛЕНИЯ С ДРУГИМ СРЕДНИМ ЗНАЧЕНИЕМ, ПОЛУЧИВШИМСЯ ПОСЛЕ ОТБОРА.

Джордж Удни Юл и метод наименьших квадратов

В конце XIX века студент Пирсона Джордж Удни Юл (Yule) (1875–1951) ввел новый подход к интерпретации корреляции и регрессии с концептуально новым использованием *метода наименьших квадратов,* который является математическим инструментом для корректировки влияния ошибок при построении кривых регрессии по точкам.

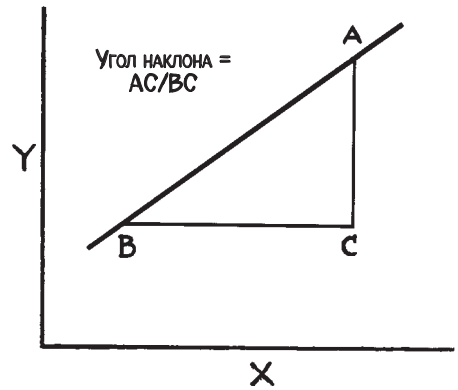


Этот метод вычисляет наиболее приближенную кривую для данных наблюдения, уменьшая сумму квадратов вертикальных отклонений от каждой точки данных к кривой регрессии.

Несмотря на то что метод наименьших квадратов может быть использован при анализе линий регрессии, бóльшая часть замешательств, связанных с регрессией к среднему значению, может быть приписана тем, кто забывает, что регрессия к среднему значению Гальтона состоит из *двух* кривых регрессии, а не одной кривой, которую можно было бы использовать для предсказания будущих исходов (с помощью метода наименьших квадратов).

Корреляция против регрессии

Несмотря на то что Гальтон хотел измерить корреляцию роста между отцами и сыновьями, в 1896 году Пирсон открыл, что метод Гальтона для поиска «со-отношения» (co-relation), как он его назвал, измеряет угол наклона линии регрессии, который является коэффициентом регрессии.



Гальтон строил линию с произвольным наклоном, а затем проверял, будет ли угол наклона равняться 1. Если значение равнялось 1, это означало, что предсказанный рост детей был аналогичным родительскому. Если значение было меньше 1, рост детей стремился к среднему значению и, как следствие, давал более умеренные значения роста.

Дилемма Гальтона

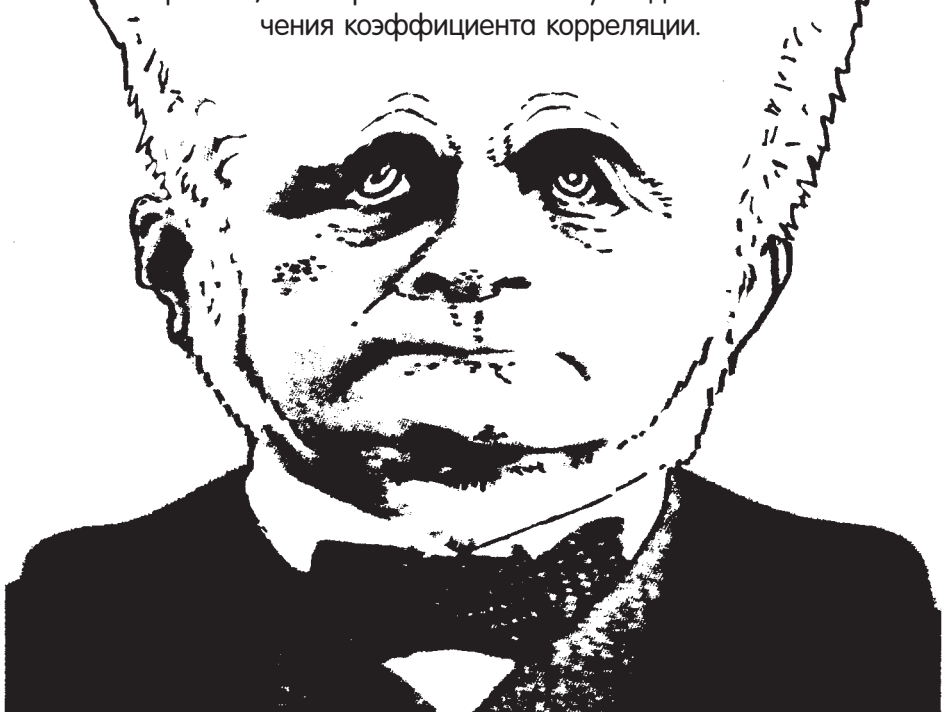
Как так получалось, что, когда Гальтон пытался найти математическую формулу для корреляции, он всегда приходил к измерению регрессии?

Пирсон прояснил работу Гальтона.

Я показал, что ошибка Гальтона была в том, что он предполагал, что есть «одинаковая колеблемость» между родителем и потомком (т. е. что изменчивость, приведенная к некоему стандарту, должна давать одинаковые числовые значения).

Пирсон смог измерить эту изменчивость отцов и сыновей по отдельности, используя свой метод среднеквадратического отклонения. Затем он показал, что, если среднеквадратические отклонения характеристики потомка и родителя имеют одинаковые числовые значения, из этого следует, что коэффициент регрессии и коэффициент корреляции также будут иметь одинаковые значения. Однако он подчеркивал, что коэффициент корреляции и коэффициент регрессии практически всегда будут все-таки разными.

Итак, Гальтон объединял воедино понятия корреляции и регрессии в своей работе. Пирсон преодолел одностороннюю концепцию Гальтона о регрессии, таким образом освободив его анализ от узости человеческой наследственности, и превратил исследования Гальтона в абсолютно статистическую концепцию. Так как Пирсон показал, что формула Гальтона измеряла регрессию, он сохранил гальтоновскую *r* для обозначения коэффициента корреляции.



Пирсоновская корреляция произведения моментов

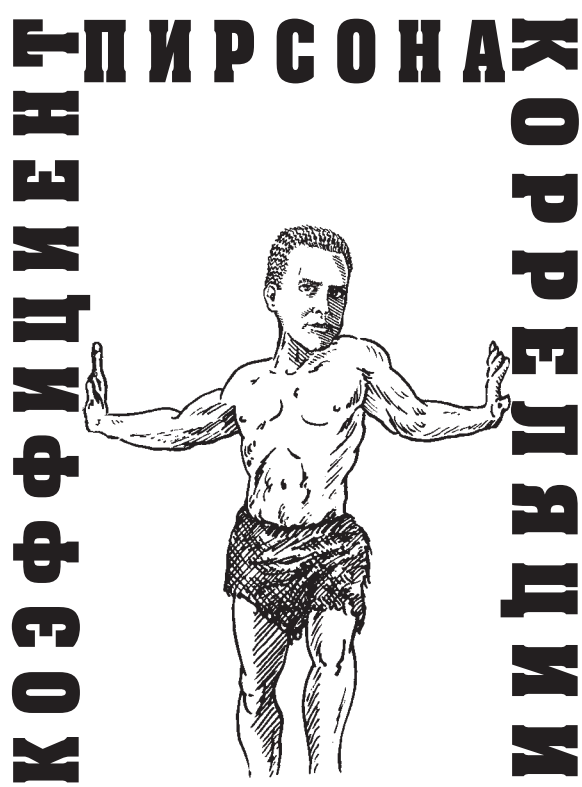
Создав системный метод моментов, Пирсон придумал строгую математическую формулу для корреляции. Он показал, что оптимальные значения угла наклона линии регрессии и коэффициент корреляции могут быть вычислены на основании произведения моментов, где *x* и *y* — это отклонения наблюдаемых значений от своих арифметических средних, соответственно. Пирсон нашел наилучшую формулу, которую в 1896 году назвал коэффициентом корреляции Пирсона (коэффициент корреляции произведения моментов):

$$r = \frac{\Sigma(xy)}{(S_x)(S_y)} = \frac{\text{ковариация}}{(\text{стандартное отклонение x}) \times (\text{стандартное отклонение y})}$$

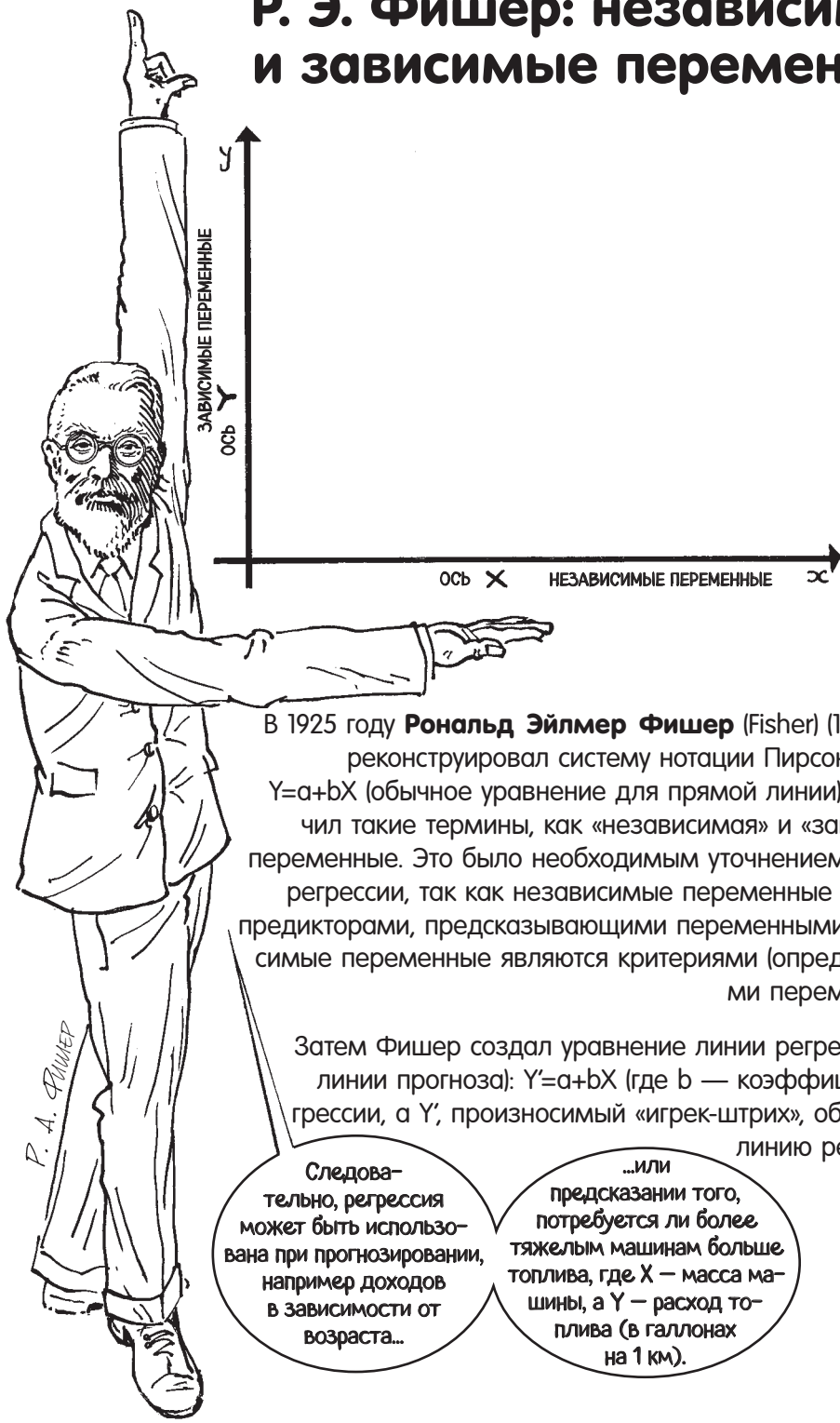
Ковариация Σ(xy) измеряет, как сильно отклонения двух случайных величин соотносятся друг с другом (см. с. 99).

Затем Пирсон определил, что коэффициент регрессии вычисляется по формуле:

$$b = \frac{\Sigma(xy)}{S^2_x} = \frac{\text{ковариация}}{\text{дисперсия x}}$$



Р. Э. Фишер: независимые и зависимые переменные



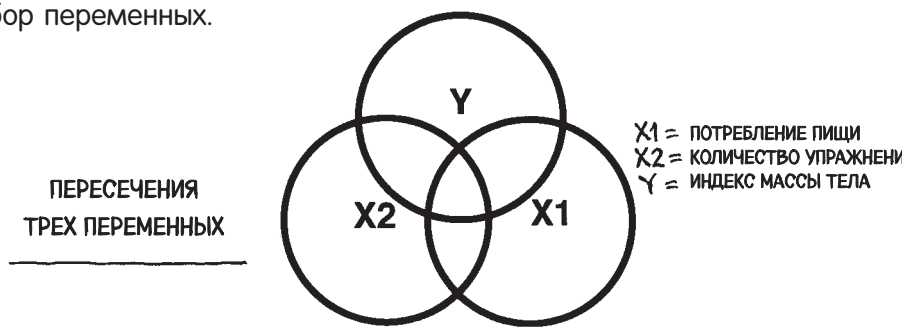
В 1925 году **Рональд Эйлмер Фишер** (Fisher) (1890–1962) реконструировал систему нотации Пирсона, введя $Y=a+bX$ (обычное уравнение для прямой линии), и включил такие термины, как «независимая» и «зависимая» переменные. Это было необходимым уточнением понятия регрессии, так как независимые переменные являются предикторами, предсказывающими переменными, а зависимые переменные являются критериями (определяющими переменными).

Затем Фишер создал уравнение линии регрессии (или линии прогноза): $Y'=a+bX$ (где b — коэффициент регрессии, а Y', произносимый «игрек-штрих», обозначает линию регрессии).

Обычная корреляция и множественная корреляция

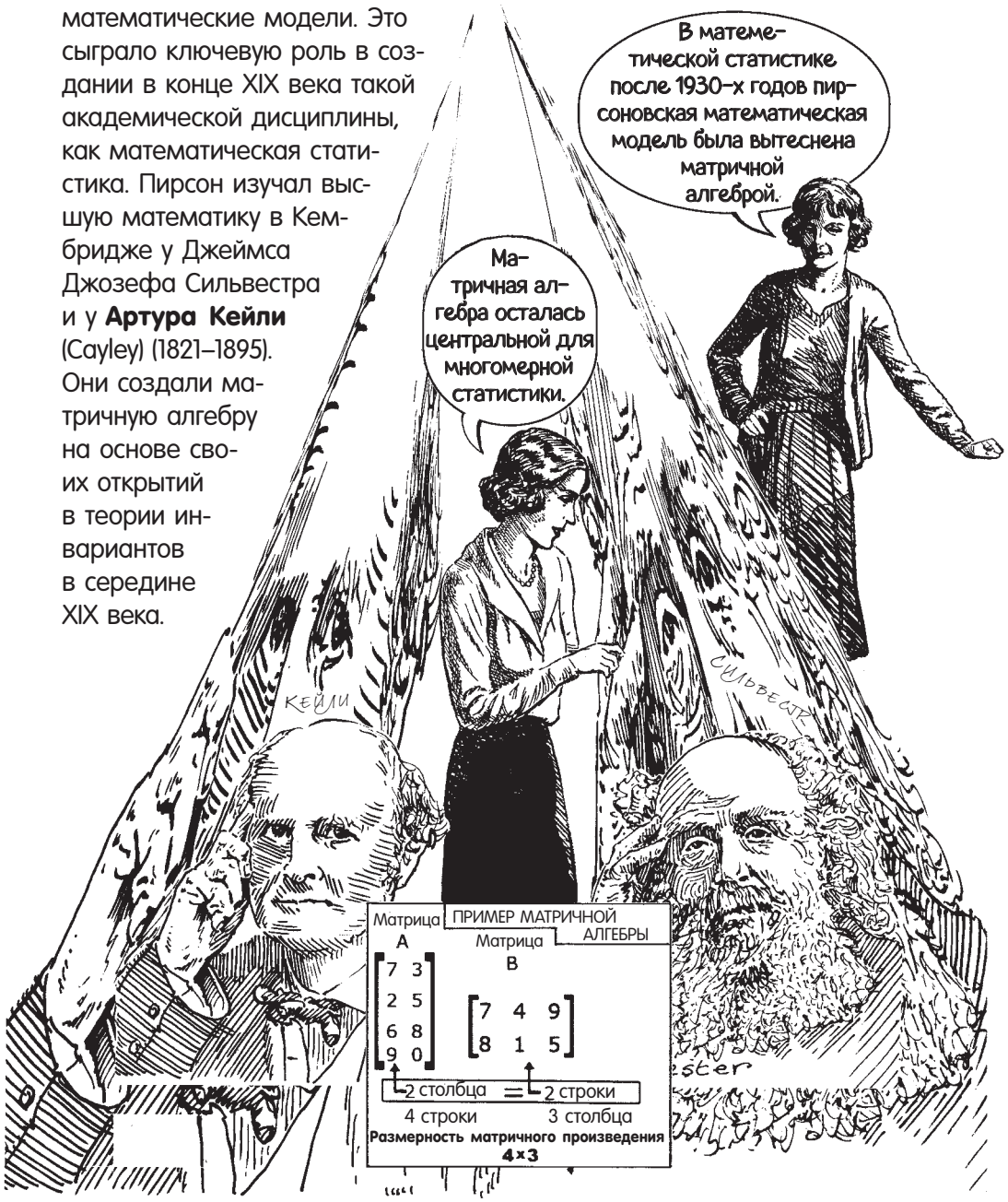
Пирсон ввел понятие **обычной корреляции**, измеряя линейные взаимосвязи между *двумя* непрерывными величинами, такими как взаимосвязь между ростом отца и ростом сына.

Эта работа стала основой развития метода **множественной регрессии**. Подобно обычной регрессии, она включает в себя линейное прогнозирование, но вместо одной прогнозируемой переменной может быть использован набор переменных.

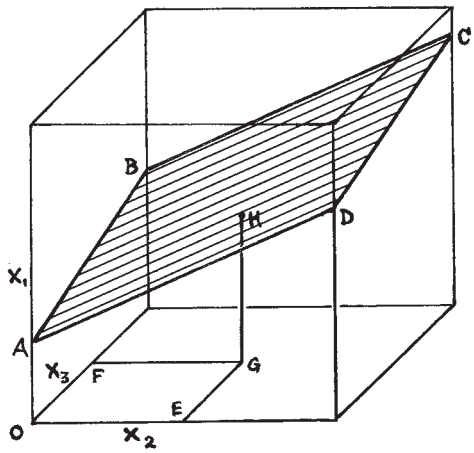


ВЫСШАЯ МАТЕМАТИКА И МАТРИЧНАЯ АЛГЕБРА

Для вычисления коэффициента множественной корреляции Пирсон использовал более сложные математические модели. Это сыграло ключевую роль в создании в конце XIX века такой академической дисциплины, как математическая статистика. Пирсон изучал высшую математику в Кембридже у Джеймса Джозефа Сильвестра и у **Артура Кейли** (Cayley) (1821–1895). Они создали матричную алгебру на основе своих открытий в теории инвариантов в середине XIX века.

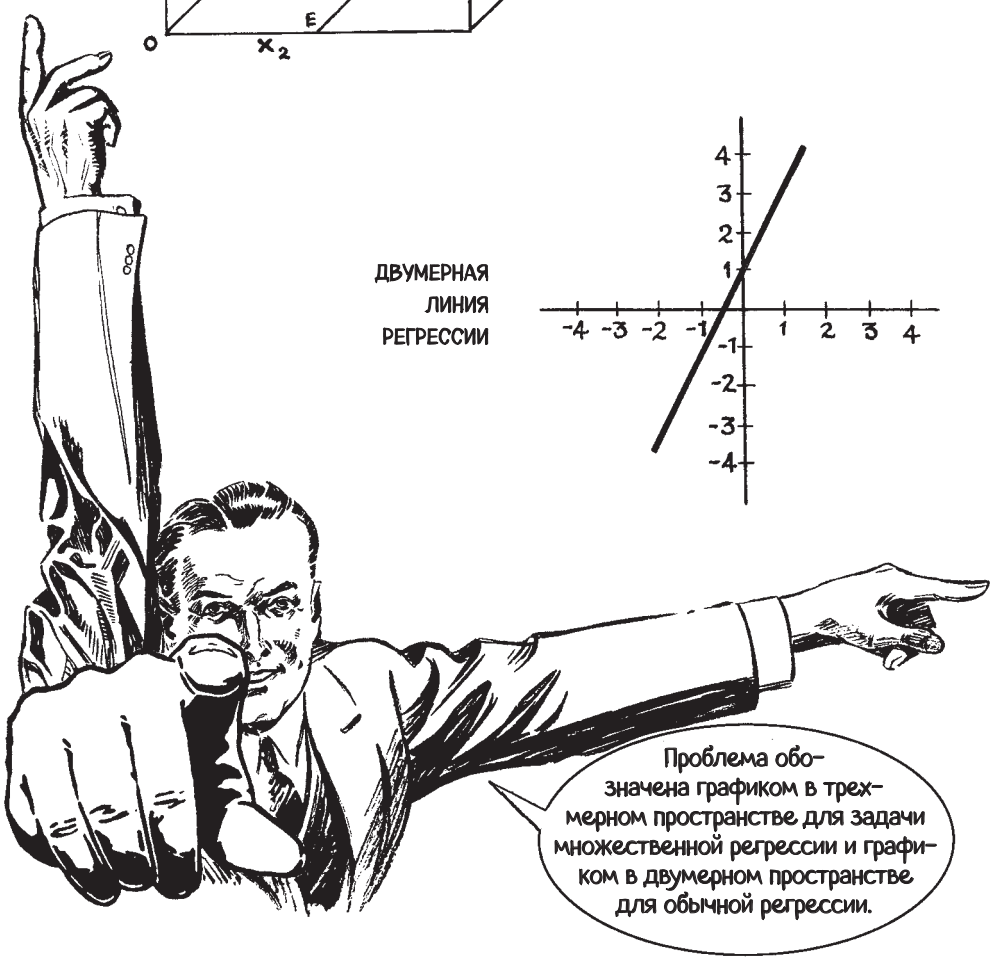


Этот новый, более высокий уровень математики позволял статистикам найти сложные математические решения для статистических проблем в многомерном (или n-мерном) пространстве, в котором двумерная модель является уже недостаточной.



ГЕОМЕТРИЧЕСКОЕ ПРЕДСТАВЛЕНИЕ МНОЖЕСТВЕННОЙ РЕГРЕССИИ (НА ПЛОСКОЙ ПОВЕРХНОСТИ)

ДВУМЕРНАЯ ЛИНИЯ РЕГРЕССИИ



Статистический контроль

Ученые используют два типа контроля во время своих исследований: экспериментальный и статистический.

В 1895 году Пирсон предложил один из способов статистического контроля некоторых переменных, введя **частную корреляцию**, которая используется только вместе с множественной корреляцией, следовательно, включает в себя три и более переменных.

Это корреляция между зависимой переменной и одной из независимых переменных. При этом исследователь убирает статистическое влияние всех прочих независимых переменных на эту искомую независимую переменную. Как следствие, исследователь может математически изолировать эту переменную, тогда как экспериментально она вообще не может быть изолирована. Статистики относятся к таким случаям так, как если бы одной из переменных просто не существовало (как мы увидим далее, частная корреляция связана с анализом ковариации Р. Э. Фишера).

Однако если бы исследователи захотели рассмотреть изолированно лишь один эффект сокращения калорий, они могли бы использовать частную корреляцию для удаления влияния переменных потребления жира и упражнений из полного набора независимых переменных. Такое исследование показало бы исключительно роль потребления калорий в снижении веса.

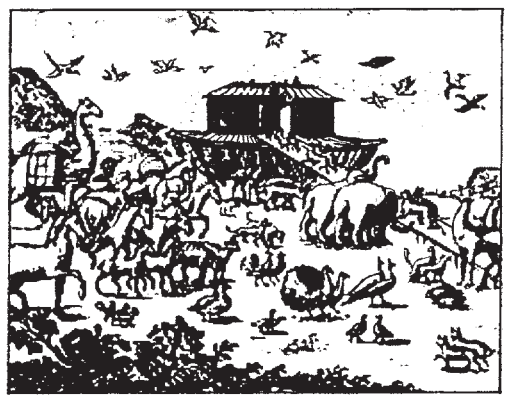
Джордж Удни Юл позже ввел **частную корреляцию**, в которой статистик элиминировал эффекты одной или нескольких независимых переменных *сразу и* для зависимой переменной, и для одной из прочих оставшихся независимых. Частная корреляция, понимаемая таким образом, помогает выявить ложную корреляцию (см. с. 118).

Дискретные взаимосвязи 2x2

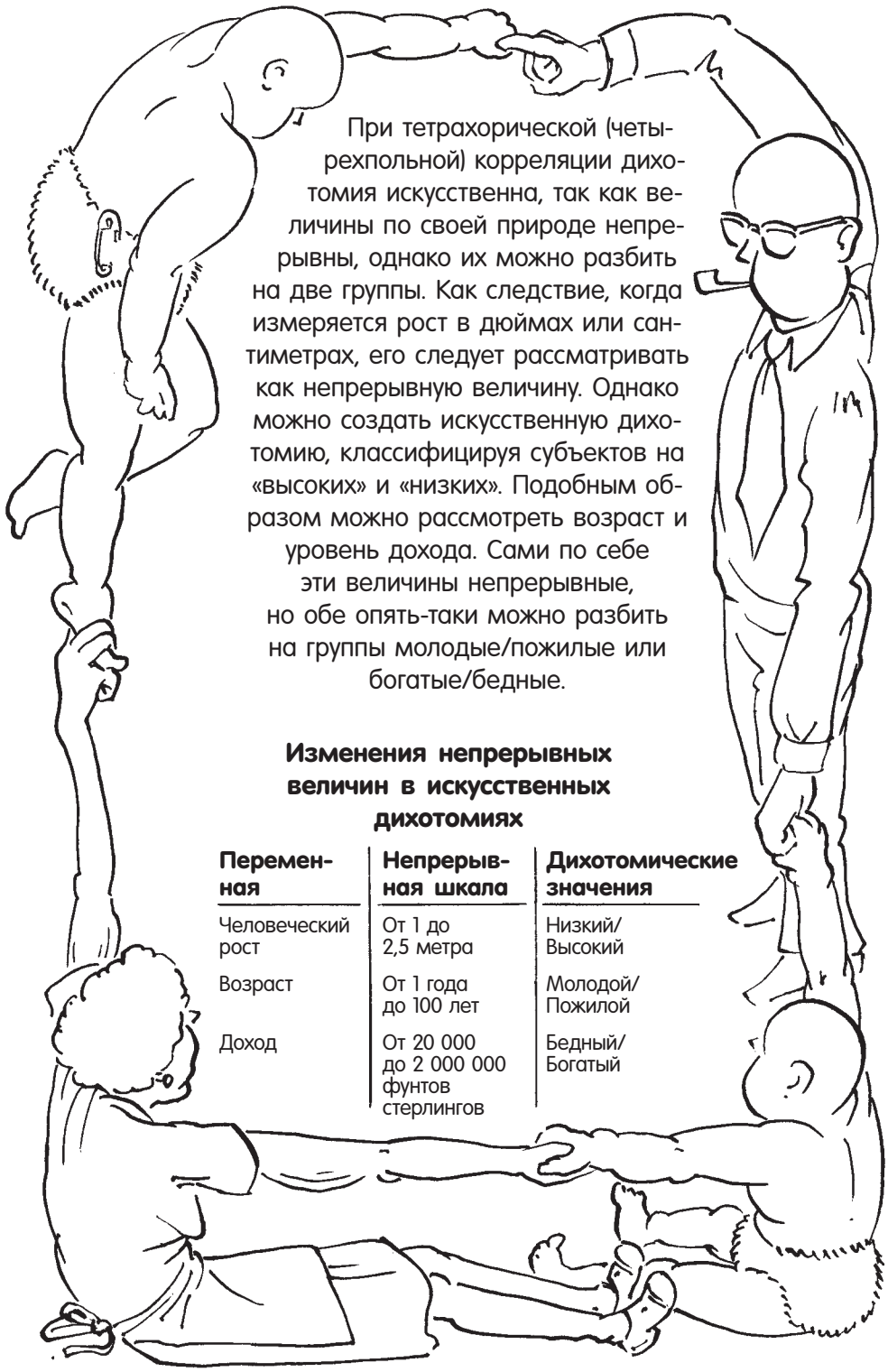
В 1900 году Пирсон ввел два новых метода: **тетрахорический** (т. е. «четырехэлементный») коэффициент корреляции (r_t); и свой **фи**-коэффициент (phi coefficient) (ф), известный позднее как «фи-коэффициент Пирсона» для дискретных величин. Оба метода измеряли ассоциативные связи между двумя переменными, размещенными в таблицах 2x2 (или четырехклеточных). Эти величины могли быть помещены в две взаимоисключающие категории (назывались «дихотомическими» переменными).

ВЫЖИВШИЕ	УМЕРШИЕ	
a	b	СБЕЖАВШИЕ
c	d	ИНЦИДЕНТ
		ЗАРАЖЕННЫЕ

ПРИМЕР ИСПОЛЬЗОВАНИЯ ЧЕТЫРЕХКЛЕТОЧНОЙ ТАБЛИЦЫ (МАТРИЦЫ) ИЗ ИССЛЕДОВАНИЯ ПИРСОНА ОТ 1904 ГОДА ОБ ЭФФЕКТИВНОСТИ ВАКЦИНЫ, ЗАЩИЩАЮЩЕЙ ОТ БРЮШНОГО ТИФА.



Фи-коэффициент Пирсона был придуман для случая двух переменных, которые находились между собой в *истинной* (подлинной) дихотомии. Как следствие, эти переменные не были непрерывными. Эта техника широко используется психометристами для создания тестов в ситуациях, в которых присутствует истинная дихотомия. Например, это так называемые «да/нет-тесты», (или тесты с ответами «истина» и «ложь»), которые используются эпидемиологами для оценки фактора риска, связанного с «присутствием» или «отсутствием» болезни (в сравнении со смертельными заболеваниями).



При тетрахорической (четырехпольной) корреляции дихотомия искусственна, так как величины по своей природе непрерывны, однако их можно разбить на две группы. Как следствие, когда измеряется рост в дюймах или сантиметрах, его следует рассматривать как непрерывную величину. Однако можно создать искусственную дихотомию, классифицируя субъектов на «высоких» и «низких». Подобным образом можно рассмотреть возраст и уровень дохода. Сами по себе эти величины непрерывные, но обе опять-таки можно разбить на группы молодые/пожилые или богатые/бедные.

Изменения непрерывных величин в искусственных дихотомиях

Переменная	Непрерывная шкала	Дихотомические значения
Человеческий рост	От 1 до 2,5 метра	Низкий/ Высокий
Возраст	От 1 года до 100 лет	Молодой/ Пожилой
Доход	От 20 000 до 2 000 000 фунтов стерлингов	Бедный/ Богатый

Q-статистика Юла

Юл предложил Q-статистику, которую он назвал в честь Кетле в 1899 году (через месяц после того, как Пирсон ввел свой фи-коэффициент и тетрахорическую корреляцию). Юл также искал способ измерения, который не был бы привязан к непрерывным величинам или не зависел бы от нормального распределения, как было в случае с корреляцией произведения моментов Пирсона.

Я обнаружил, что мое Q (диапазон значений от −1,00 до 1,00) оказывалось всегда чуть выше, чем тетрахорическая корреляция Пирсона.

$$Q = \frac{ad - bc}{ad + bc}$$

Социологи были первыми, кто применил в своих работах Q-статистику Юла. Ее использовали медицинские статистики в конце XX века, и она стала мерой ассоциативной связи в случаях, которые возникали напрямую из ячеек таблицы 2x2. Теперь это соотношение известно как коэффициент несогласия*, который основывался на Q-статистике Юла.

* Коэффициент несогласия (odds ratio) — это способ сравнения вероятностей наступления некоторого события, произошедшего в обоих группах (будут ли эти вероятности одинаковыми). — *Прим. науч. ред.*

Бисериальные корреляции

Пирсон придумал **бисериальную корреляцию** в 1909 году. Она относится к корреляции произведения моментов (в которой обе переменные являются непрерывными), но с одним отличием.

Как мы увидим далее, бисериальная корреляция похожа на t-статистику Стьюдента и дисперсионный анализ Фишера.

Точечно-бисериальная корреляция связана с бисериальной корреляцией Пирсона, однако здесь одна переменная непрерывная, а другая измеряется по дихотомической шкале («истинная дихотомия»), например мужской/женский пол. Этот тип корреляции представлял бы собой приближенную оценку корреляции произведения моментов, если бы в корреляции произведения моментов дихотомическая переменная была бы заменена непрерывной переменной.

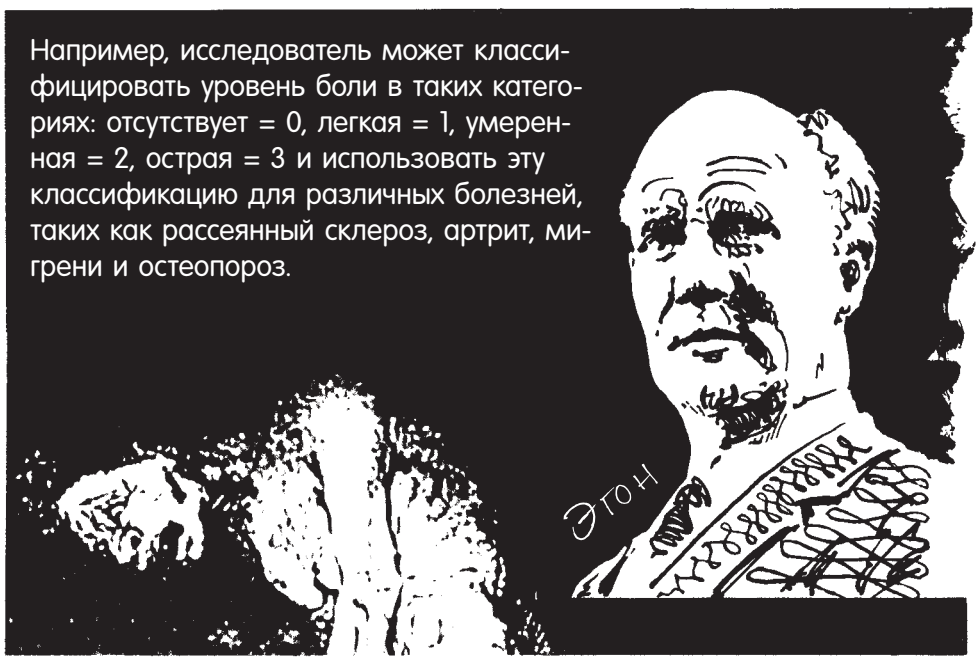
Есть два основных метода, которые широко используются психометристами для анализа исследуемых показателей при создании различных тестов на уровень интеллекта и способностей. Бисериальная корреляция обычно используется для определения корреляции между баллами исследуемого показателя и общими баллами за прохождение теста.

Трисериальная корреляция Пирсона похожа на бисериальную, в которой одна переменная непрерывная, а вторая является трихотомией (например, низкий, средний, высокий).

Эгон Пирсон и полихорические корреляции

В 1922 году Пирсон со своим сыном Эгоном придумал **полихорическую** корреляцию.

Она похожа на тетрахорическую корреляцию, за исключением того, что есть *три или более* возможных значения, которые может принимать переменная. Если тетрахорическая корреляция ограничена таблицей сопряжения признаков 2x2, где переменные могут принимать только бинарные значения (т. е. 0, 1), то для полихорической корреляции используется таблица nxn, а значения переменных полисериальны (0, 1, 2, 3, 4…). Как следствие, в таблице содержится три и более категорий.



УРОВНИ БОЛИ

Вид заболевания	Отсутствует = 0	Легкая = 1	Умеренная = 2	Острая = 3
Рассеянный склероз				
Мигрени				
Артрит				
Остеопороз				

Ранговая корреляция изучает взаимосвязи между различными порядковыми номерами (рангами) одних и тех же данных. Она позволяет измерить зависимость между двумя порядковыми номерами, и оценить ее статистическую значимость. Два основных метода были придуманы студентом Карла Пирсона **Чарльзом Спирменом** (Spearman) (1863–1945) и **Морисом Кендаллом** (Kendall). Три других теста — это критерий знаковых рангов Уилкоксона (Wilcoxon), U-критерий Манна — Уитни (Mann — Whitney) и анализ рангов Крускала — Уоллиса (Kruskal — Wallis).

В принципе этот метод — один из особых случаев коэффициента корреляции произведения моментов Пирсона, в котором данные преобразованы (еще до вычисления коэффициента) в ранги, от высшего к низшему.

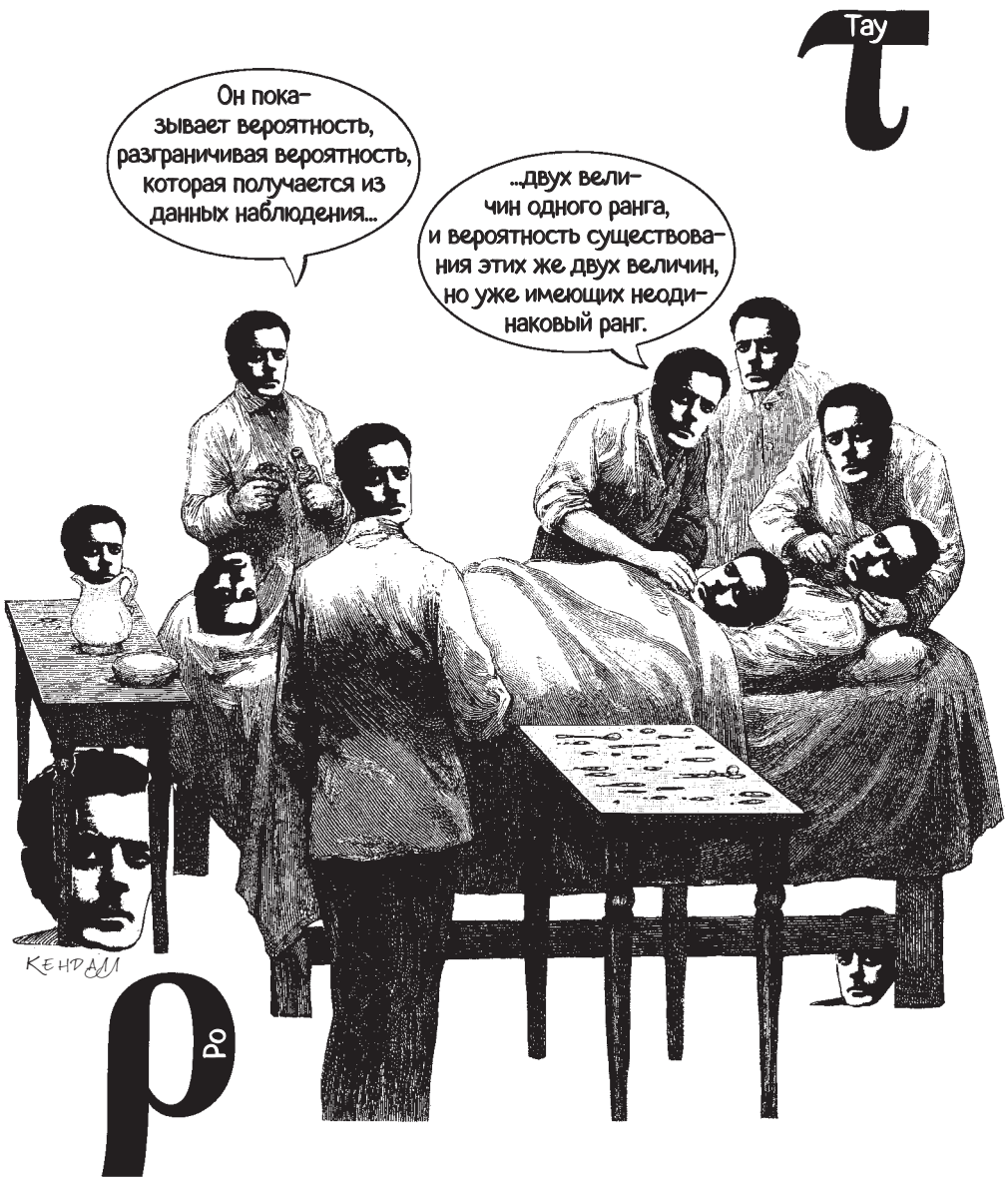
Факторный анализ

Спирмен также интересовался идеями Гальтона об измерении индивидуальных различий человеческих способностей и его ранними подходами к измерению интеллекта. Используя корреляцию произведения моментов Пирсона и метод главных компонент*, который Пирсон придумал в 1901 году, Спирмен создал новый статистический метод, известный как **факторный анализ**, который сводил набор сложных данных в более удобную для работы форму, позволяющую увидеть *структуру взаимосвязи* между переменными.

* Общая статистическая процедура по поиску оптимального представления набора данных, обладающих свойством коррелированности друг с другом. — *Прим. науч. ред.*

Тау-коэффициент Мориса Кендалла

В 1938 году английский статистик **Морис Кендалл** (1907–1983) создал другой метод для ранговой корреляции, известный как метод Кендалла. Этот метод — схема, основанная на согласованности или несогласованности упорядоченных (ранговых) данных.



Тау-коэффициент Кендалла часто используется в выборках, которые шире, чем те, с которыми работает метод Спирмена и его коэффициент «ро» (ρ).

Корреляция против ассоциативной связи

Эти термины используются для описания двух разных процедур, измеряющих статистические взаимосвязи.

I. Методы **корреляции** — это:

Пирсоновская обычная, множественная и частная корреляция

Бисериальная корреляция

Трисериальная (и полисериальная) корреляция

Тетрахорическая корреляция

и частная корреляция Юла

II. Измерение **ассоциативной связи**, в которой обе переменные являются номинальными:

ФИ-КОЭФФИЦИЕНТ

Статистика хи-квадрат (см. с. 153–156)

и Q-статистика Юла

III. **Смешанные (прикладные) измерения**, в которых одна переменная дискретная, а вторая непрерывная:

Полихорическая корреляция

Тау-коэффициент Кендалла

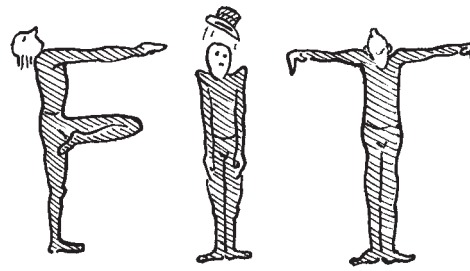
коэффициент «ро» (ρ)

Критерий знаковых рангов Уилкоксона

U-критерий Манна — Уитни

и анализ рангов Крускала — Уоллиса

Статистические критерии согласия

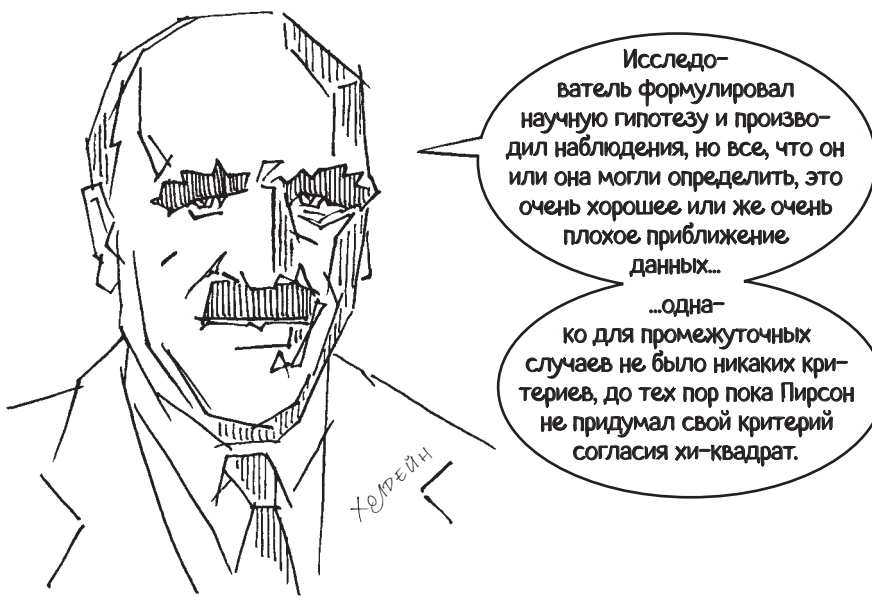


Одним из способов использования нормального распределения для анализа или интерпретации данных, является метод, называемый *статистическим критерием согласия*. Он позволяет ученому-статистику увидеть, насколько точно данные соответствуют нормальному распределению.

Наш поборник нормального распределения, Адольф Кетле сделал одну из первых попыток построить кривую нормального распределения по данным наблюдений в 1840 году, которую Гальтон начал использовать в 1863 году. Методы Кетле были графическими, для построения он использовал таблицу данных, основанную на биномиальном распределении, вместо того чтобы использовать приближения кривой нормального распределения. Бóльшая часть трудов Гальтона не состояла из построения кривых в чистом виде, наоборот, он сравнивал вычисленные им значения с таблицей нормального распределения вероятностей.

В 1877 году Вильгельм Лексис придумал соотношение Лексиса L как статистический критерий согласия для определения того, согласуется ли эмпирическое распределение с нормальным распределением. В 1887 году Фрэнсис Исидро Эджуорт придумал статистический критерий, согласия который был основан на приближении нормального закона к биномиальному распределению. Хотя многие другие ученые XIX века пытались придумать статистические критерии согласия, им (в отличие от Пирсона) не удалось создать теоретическую базу для своих формул.

До того как Пирсон придумал новый статистический критерий согласия, обычный метод состоял в сравнении ошибок наблюдения с таблицей распределения вероятностей, основанной на кривой нормального распределения, или, если использовать графические средства, в сравнении с диаграммой плотности распределения. Как это объяснял в 1936 году биолог-эволюционист **Джон Бердон Сандерсон Холдейн** (Haldane) (1892–1962):

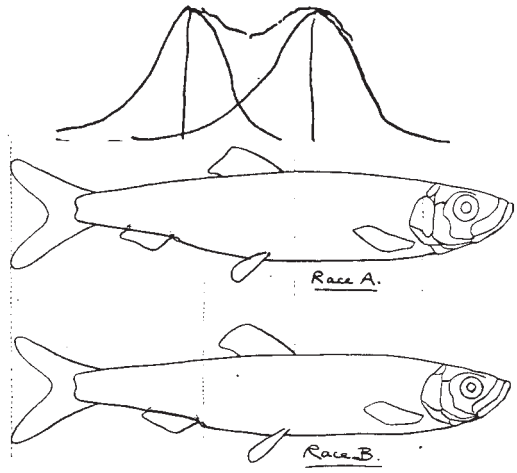


Подбор эмпирических кривых для асимметричных распределений

Интерес Пирсона в области подбора эмпирических кривых был инспирирован работой Велдона о зеленых крабах Плимута. Когда в 1892 году Велдон открыл, что часть данных о крабах не подходит под единую кривую нормального распределения, а вместо этого получаются две кривые, — эффект, который он назвал «двухвершинным», или бимодальным, распределением, — Велдон обратился к Пирсону за помощью.

Пирсон искал другой способ интерпретации данных, который не был бы попыткой свести их к кривой нормального распределения, как это прежде делали Кетле и Гальтон. Пирсон и Велдон считали, что важно придать смысл форме кривой, не искажая ее, так как это может помочь открыть что-то новое об исследуемых видах.

Система хи-квадрат



Outlines traced from Heincke's figures of two typical Kiel Herring, one belonging to his Race A, the other to his Race B.

РИСУНКИ ВЕЛДОНА О РАЗЛИЧИЯХ СЕЛЬДИ, КОГДА ОН И ПИРСОН ИСКАЛИ ОТЛИЧИТЕЛЬНЫЕ ПРИЗНАКИ ОСОБЕЙ

Непрерывная работа Пирсона над подбором эмпирических кривых на протяжении 1890-х годов вела к тому, что ему был нужен критерий для определения того, насколько точно подбирается кривая. Это привело его к формулировке различных статистических критериев согласия. К концу 1896 года Пирсон работал над созданием статистического критерия согласия для асимметричных распределений, нужных биологам и экономистам. Результатом этой работы стал критерий согласия хи-квадрат, который был создан в 1900 году.

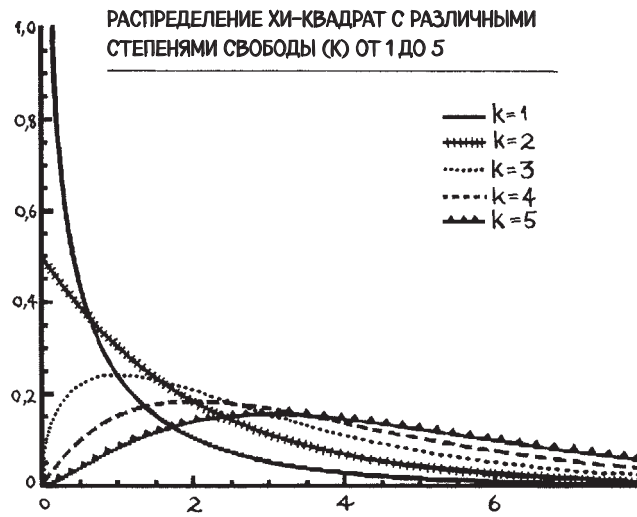
Есть три основных элемента системы хи-квадрат (χ^2) Пирсона:

1. Распределение вероятностей хи-квадрат, опубликованное в 1900 году.
2. Статистический критерий согласия, придуманный в 1900 году.
3. Критерий согласия хи-квадрат (взаимосвязи) для факторных таблиц, придуманный в 1904 году (затем переименованный в «статистику хи-квадрат» в 1923 году Р. Э. Фишером).

Однако что же так отличало распределение хи-квадрат и критерий согласия хи-квадрат от всех других распределений и критериев?

Если нормальное распределение используется обычно для непрерывных данных, которые дают симметричную, колоколообразную кривую, то распределение хи-квадрат можно применять к дискретным данным, имеющим различные распределения, такие как асимметричное, биномиальное или распределение Пуассона.

Критерий хи-квадрат Пирсона основывается на двух различных гипотезах: статистический критерий согласия определяет, насколько точно эмпирическое распределение, основанное на данных наблюдения или эксперимента, может эффективно описывать выборку, взятую из рассматриваемой генеральной совокупности (т. е. насколько точно экспериментальные данные соотносятся с теоретическим распределением хи-квадрат).



Наоборот, коэффициент сопряженности (признаков) хи-квадрат, который измеряет взаимосвязь, проверяет разницу между наблюдаемыми значениями и теоретически ожидаемыми значениями факторной таблицы.

В приведенном ниже примере политолог-аналитик хочет определить, женщины или мужчины более склонны голосовать за республиканцев или демократов на американских президентских выборах.

Предпочтения голосования в таблице 2x2:

Политическая партия	Пол		Итого
	Женский	Мужской	
Демократическая	a	b	a+b
Республиканская	c	d	c+d
Итого	**a+c**	**b+d**	**N**

Статистика хи-квадрат для факторных таблиц (таблиц сопряженности признаков) наилучшим образом иллюстрируется вычислительной формулой, которую придумал Пирсон в 1904 году для факторных таблиц 2x2:

$$\chi^2 = \Sigma \frac{n\,(ad - bc)^2}{(a + b)\,(c + d)\,(b + d)\,(a + c)}\,.$$

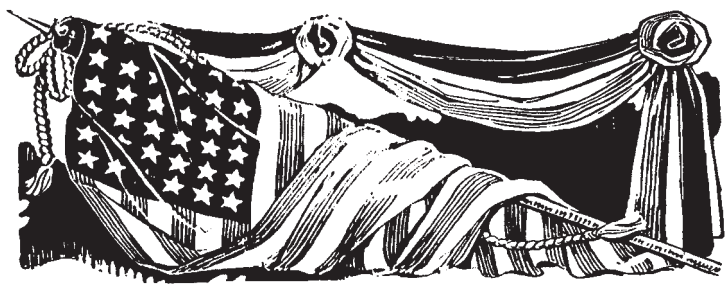
Статистика хи-квадрат может показать, что женщины более склонны голосовать за демократов, а мужчины — за республиканцев.

Несмотря на то, что два статистических критерия согласия хи-квадрат выполняют разные функции, они могут быть выражены математически (в современных терминах) как:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

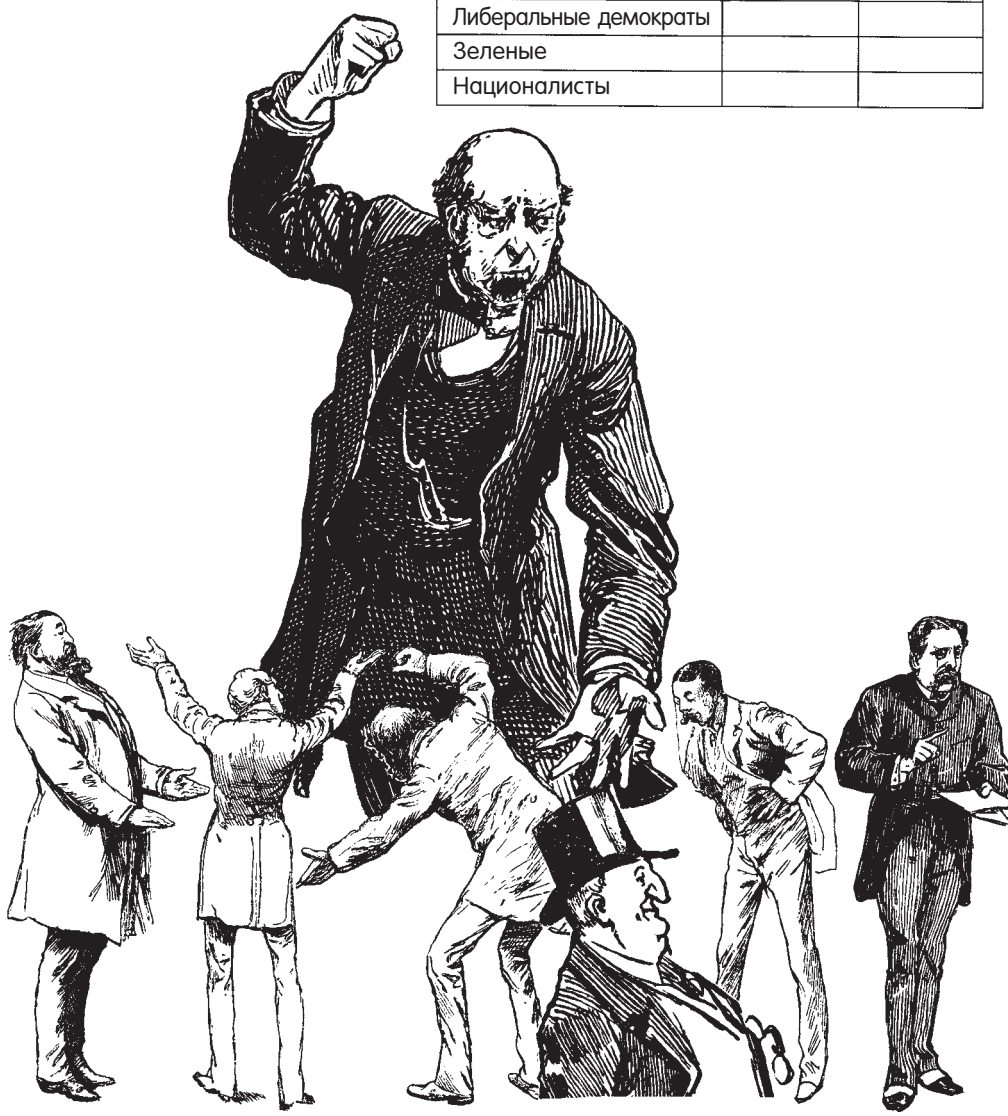
или

$$\text{хи-квадрат} = \text{сумма всех значений} \left(\frac{(\text{наблюдаемое число} - \text{ожидаемое число})^2}{\text{ожидаемое число}} \right)$$



Статистика хи-квадрат гибкая и позволяет работать более чем с одной категорией, но тогда следует использовать более общую формулу. Как следствие, если политолог-аналитик захочет узнать о британских выборах в парламент, в которых участвуют более чем две партии, и о том, за какие партии склонны голосовать мужчины и женщины, то наблюдаемые значения будут рассматриваться в таблице 2x5.

Политическая партия	**Пол**	
	Женский	**Мужской**
Лейбористы		
Консерваторы		
Либеральные демократы		
Зеленые		
Националисты		



Интерпретация результатов по степеням свободы

В отличие от корреляции, в случае которой Пирсон мог взглянуть на числа (например, 0,90, 0,50 или 0,21) и понять, что они обозначают высокую, среднюю или низкую корреляцию, в хи-квадрат статистике такое невозможно. Исследователь не может взглянуть на значение, получившееся из формулы, и понять, что оно означает, не применив дополнительных методов.

Для интерпретации вычисленных значений хи-квадрат Пирсон придумал так называемый поправочный коэффициент. В 1922 году Р. Э. Фишер сформулировал понятие степеней свободы для определения значимости результатов, полученных при использовании метода хи-квадрат. Степени свободы основаны на количестве наблюдений в выборке и универсальным образом используются во многих статистических методах.

Таблица вероятностей хи-квадрат

В 1900 году Пирсон и его студентка Элис Ли (Lee) (1858–1939) создали таблицу вероятностей хи-квадрат. Год спустя другой студент, **Уильям Полин Эльдертон** (Elderton) (1877–1962), усовершенствовал ее. Иметь доступ к таблице вероятностей означает, что исследователь мог сверить вычисленные значения хи-квадрат и необходимый поправочный коэффициент для определения того, являются ли результаты статистически значимыми или нет.

Хотя Эджуорт обсуждал статистические критерии значимости уже в 1885 году, критерий хи-квадрат Пирсона сделал возможным определение статистической значимости результатов в совершенно новом масштабе, который был недосягаем до него. Последующие поколения статистиков показали, что были и другие факторы, которые влияли на истинные степени свободы для критерия хи-квадрат.

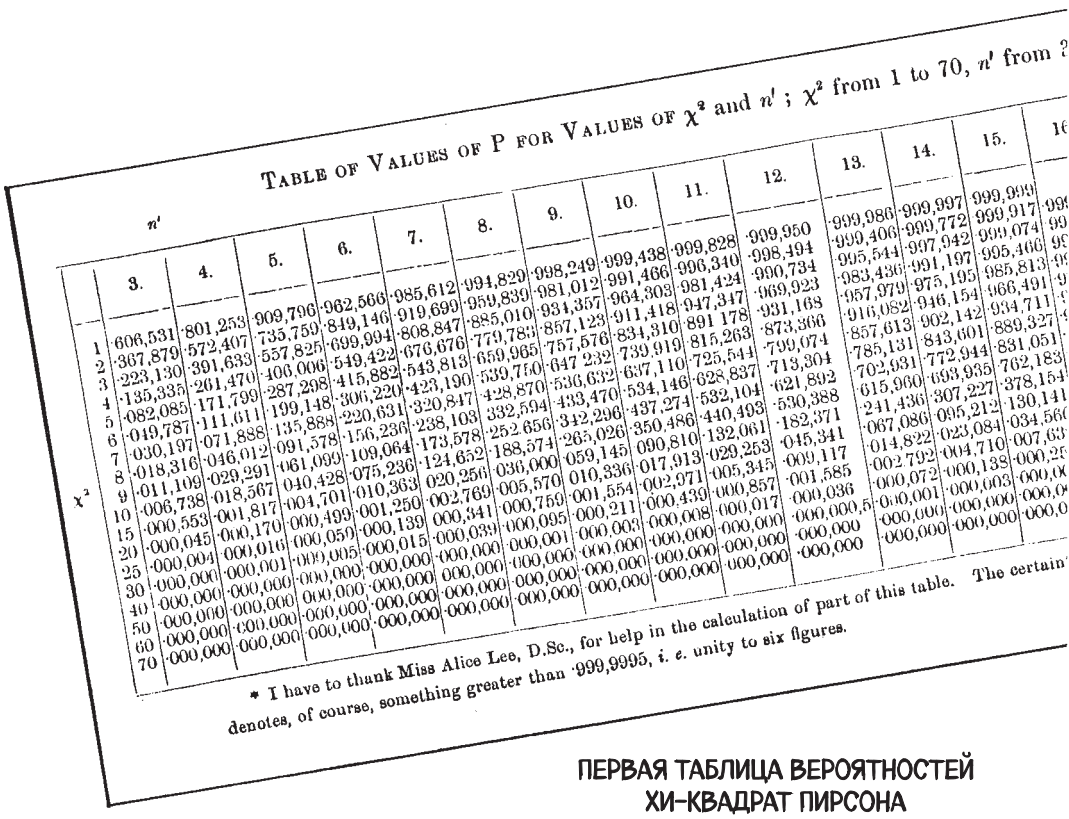


TABLE OF VALUES OF P FOR VALUES OF χ^2 and n'; χ^2 from 1 to 70, n' from 2

χ^2 \ n'	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16
1	·606,531	·801,253	·909,796	·962,566	·985,612	·994,829	·998,249	·999,438	·999,828	·999,950	·999,986	·999,997	·999,999	·99
2	·367,879	·572,407	·735,759	·849,146	·919,699	·959,839	·981,012	·991,466	·996,340	·998,494	·999,406	·999,772	·999,917	·99
3	·223,130	·391,633	·557,825	·699,994	·808,847	·885,010	·934,357	·964,303	·981,424	·990,734	·995,544	·997,942	·999,074	·99
4	·135,335	·261,470	·406,006	·549,422	·676,676	·779,783	·857,123	·911,418	·947,347	·969,923	·983,436	·991,197	·995,466	·99
5	·082,085	·171,799	·287,298	·415,882	·543,813	·659,965	·757,576	·834,310	·891 178	·931,168	·957,979	·975,195	·985,813	·9
6	·049,787	·111,611	·199,148	·306,220	·423,190	·539,750	·647 232	·739,919	·815,263	·873,366	·916,082	·946,154	·966,491	·9
7	·030,197	·071,888	·135,888	·220,631	·320,847	·428,870	·536,632	·637,110	·725,544	·799,074	·857,613	·902,142	·934,711	·
8	·018,316	·046,012	·091,578	·156,236	·238,103	·332,594	·433,470	·534,146	·628,837	·713,304	·785,131	·843,601	·889,327	·
9	·011,109	·029,291	·061,099	·109,064	·173,578	·252 656	·342,296	·437,274	·532,104	·621,892	·702,931	·772,944	·831,051	·
10	·006,738	·018,567	·040,428	·075,236	·124,652	·188,574	·265,026	·350,486	·440,493	·530,388	·615,960	·693,935	·762,183	·
15	·000,553	·001,817	·004,701	·010,363	·020,256	·036,000	·059,145	·090,810	·132,061	·182,371	·241,436	·307,227	·378,154	·
20	·000,045	·000,170	·000,499	·001,250	·002,769	·005,570	·010,336	·017,913	·029,253	·045,341	·067,086	·095,212	·130,141	·
25	·000,004	·000,016	·000,059	·000,139	·000,341	·000,759	·001,554	·002,971	·005,345	·009,117	·014,822	·023,084	·034,566	·
30	·000,000	·000,001	·000,005	·000,015	·000,039	·000,095	·000,211	·000,439	·000,857	·001,585	·002,792	·004,710	·007,63	·
40	·000,000	·000,000	·000,000	·000,000	·000,001	·000,003	·000,008	·000,017	·000,036	·000,072	·000,138	·000,26	·
50	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,001	·000,003	·000,0	·
60	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,0	·
70	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,000	·000,0	·

* I have to thank Miss Alice Lee, D.Sc., for help in the calculation of part of this table. The certain
denotes, of course, something greater than ·999,9995, *i. e.* unity to six figures.

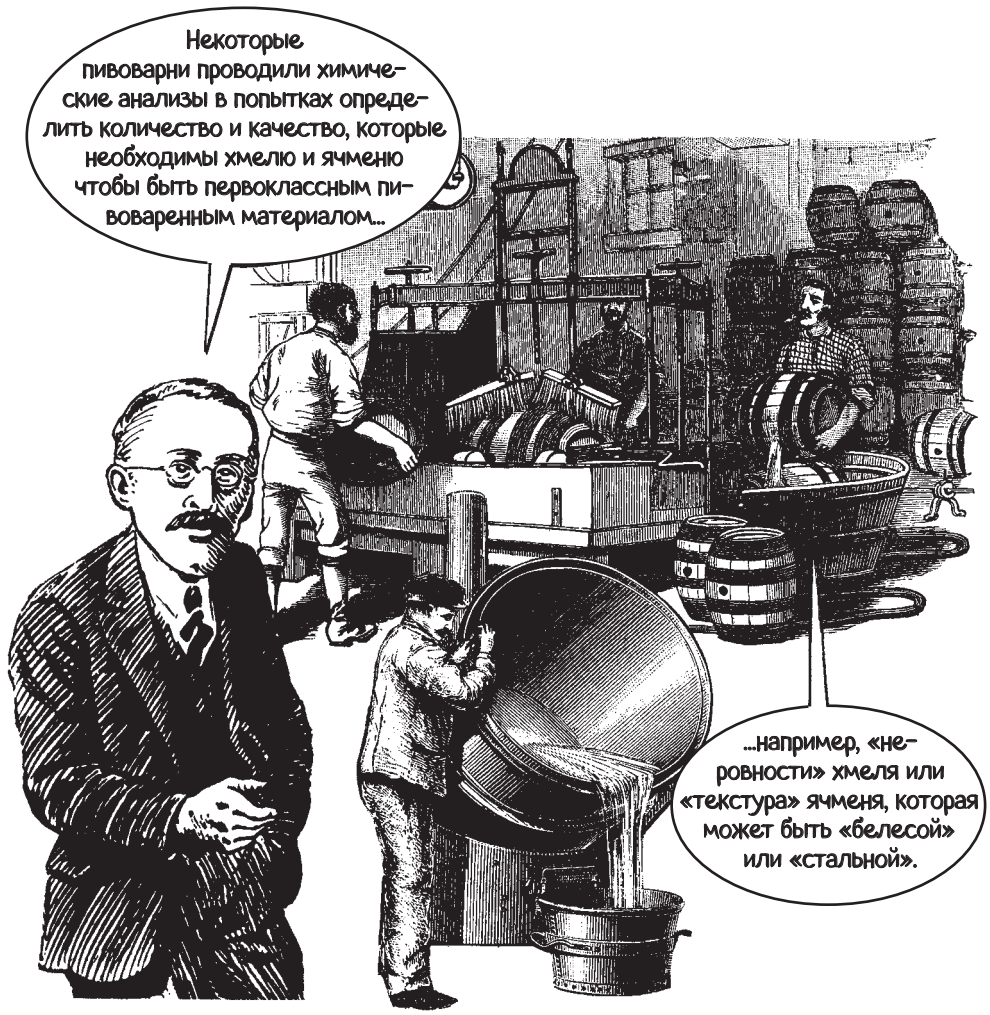
ПЕРВАЯ ТАБЛИЦА ВЕРОЯТНОСТЕЙ ХИ-КВАДРАТ ПИРСОНА 1900 ГОДА

Статистический критерий для пивоварни Guinness

Первый статистический критерий для определения контроля качества на производстве был придуман химиком и статистиком Уильямом Сили Госсетом (Gosset), который работал на пивоварне Guinness в начале XX века. Так как Госсет был связан обязательством с Guinness не публиковать работы под своим именем (возможно, оттого, что в Guinness не хотели, чтобы конкуренты узнали о статистических наработках для производства и о приглашенных на работу профессиональных статистиках), он взял псевдоним «Стьюдент». Это была стандартная практика в Guinness: ассистент Госсета, Эдвард М. Сомерфилд (Somerfield), взял псевдоним «Аламнус» (Alumnus) [букв. «выпускник»] для своих публикаций.

Определение количества пивоваренного материала

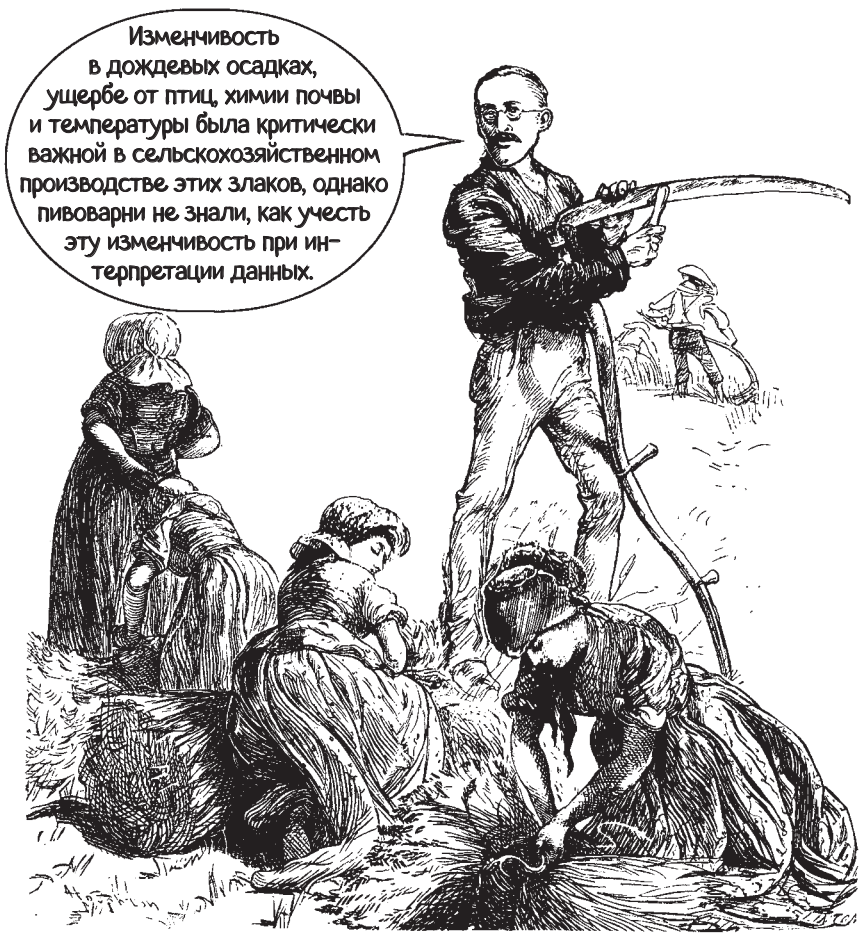
У Guinness были широкие интересы в сельском хозяйстве, в особенности в выращивании ячменя для пива. Это привело Госсета к сельскохозяйственным экспериментам и лабораторным опытам.



Однако эти качественные характеристики было сложно измерять, поэтому в Guinness не могли знать точно, что именно делало их стаут (крепкое пиво) таким популярным и как можно было поддерживать или улучшить качество. Они хотели знать, какие условия необходимы для производства тех сортов ячменя и хмеля, которые давали бы наилучшие пивоваренные качества.

Изменчивость в сельском хозяйстве

Когда Госсет устроился на работу в Guinness, он столкнулся с огромным количеством химических данных в пивоваренном производстве и заинтересовался, можно ли на их основе увидеть взаимосвязь качества исходного сырья, такого как ячмень и хмель, и качества готового продукта. Две проблемы с которыми столкнулся Госсет, когда он начал планировать свой статистический анализ, были связаны с тем, что изменчивость была высокой, а статистических наблюдений было мало.



Как следствие, Guinness нужен был способ определения того, какую изменчивость можно было проигнорировать, а какую — крайне важно учесть. Один путь анализа изменчивости — это использование статистических методов Пирсона. Госсет договорился о встрече с Пирсоном 12 июля 1905 года в Ист-Илси в Беркшире, где Пирсон проводил свой летний отпуск, находясь на доступном велосипедном расстоянии от Велдона в Оксфорде.

Малые и большие выборки

Госсет рассказал Пирсону, что одной из его главных проблем были малые размеры выборки — у него была выборка из десяти элементов для каждого сорта ячменя. По сравнению с выборками, с которыми имел дело Пирсон, это были очень малые выборки. Такая проблема привела Госсета к созданию первого статистического критерия контроля качества.

Госсет адаптировал методы Пирсона для малых выборок, а также заимствовал некоторые статистические методы, используемые в астрономии. Эти комбинированные линейные уравнения наблюдений были предназначены, однако, для очень ограниченного использования, так как имели дело с наблюдениями, производимыми в стационарных условиях. Напротив, сельскохозяйственные условия данных пивоварения были *неустойчивы*: они сильно изменялись и были подвержены влиянию изменений, сделанных в ходе лабораторных экспериментов.

Оценивая статистическую разницу между двумя среднеарифметическими

Используя астрономические методы в сочетании с методами Пирсона, Госсет создал статистический инструментарий, необходимый для его экспериментальных данных. Он хотел узнать, существовала ли принципиальная разница (со статистической точки зрения) в использовании двух типов удобрений для двух сортов ячменя, выращенных на соседних полях, при разном типе почвы, навоза и погодных условий.

Механизм вычисления разницы между двумя средними в группах (внутри каждой из групп), а также попытка найти полноценный способ интерпретации данных для случая малых выборок ранее в 1850-х годах, уже были предложены французским физиком Пьером Луи (Louis) и немецким физиком Густавом Радике (Radicke). Однако они не считались особенно успешными. Госсет придумал **z-соотношение** (или Z-критерий) для определения статистической значимости разницы между выборочным средним и средним по генеральной совокупности.

Статистические результаты для Guinness

Когда Госсет проанализировал данные по ячменю, использовав придуманный им z-критерий, он обнаружил, что лучший ячмень для Guinness — это сорт ячменя «арчер» (Archer). Как только в Guinness узнали о результатах исследования, они захотели выращивать этот сорт по всей Ирландии.

Z-соотношение Стьюдента стало первым статистическим критерием, использованным на производстве для контроля качества. Идеи Госсета, которые показали важность определения контроля качества товара, повлияли на новое поколение статистиков, включая Р. Э. Фишера, Уолтера Шухарта (Shewhart) (1891–1967) и Уильяма Эдвардса Деминга (Deming) (1900–1993).

t-Критерий Стьюдента

Фишер был так впечатлен статистическим критерием Госсета, что в 1924 году он усовершенствовал Z-критерий Госсета и, повторно введя в его науку, назвал «t-критерием Стьюдента». Фишер вычислил значения Госсета из z-таблицы и заменил ее t-таблицей, которую обозначил как «t-распределение Стьюдента». t-критерий Стьюдента теперь выражался так:

$$t = \frac{\text{выборочное среднее группы 1} - \text{выборочное среднее группы 2}}{\text{среднеквадратическая ошибка разницы}} \left\{ \frac{\bar{X}_1 - \bar{X}_2}{se_{\bar{}}} \right\}$$

среднеквадратическая ошибка

Есть три разных способа использования t-критерия:

В дальнейшем Фишер развил идеи Госсета, придумав свой «дисперсионный анализ» для своих классических экспериментов с пшеницей, проведенных на Ротамстедской экспериментальной станции в Харпендене, Хартфордшир (к северу от Лондона).

Новая эра статистики: Ротамстедские сельскохозяйственные данные

Несмотря на то, что в 1919 году Пирсон предложил Фишеру должность в Университетском колледже Лондона, Фишер принял предложение сэра Джона Расселла (Russell) по работе в Ротамстедской экспериментальной станции для анализа сельскохозяйственных данных Броудболк (Broadbalk), в ходе которого его статистические инновации принесли свои плоды.

Ротамстед — это один из наиболее древних сельскохозяйственных центров в мире, основанный в 1834 году **Джоном Беннетом Лоусом** (Lawes) (1814–1902), чьи предки владели этой землей с 1623 года.

В 1834 году химик **Джозеф Генри Гилберт** (Gilbert) (1817–1901) присоединился к Лоусу в его работе над экспериментальной культивацией в полях Броудболк. На основе своей работы они опубликовали все статистические детали своих наблюдений и экспериментов и обнаружили, что непрерывно удобряемые поля давали от 12 до 13 бушелей в год, в то время как хорошоу навоженные поля давали от 30 до 40 бушелей в год.

С концом Первой мировой войны в 1918 году связано расширение и перестройка Ротамстеда. В следующем году сельскохозяйственный химик **Эдвард Джон Расселл** (1872–1965) взял на работу кембриджского математика Фишера.

СБОР УРОЖАЯ НА ПОЛЯХ БРОУДБОЛК

Дисперсионный анализ Фишера

С 1919 по 1926 год Фишер закладывал принципы проведения экспериментов и развивал свою статистическую методологию дисперсионного анализа, которым он начал заниматься в 1916 году (ANOVA). Пока все эксперименты были завязаны на отношения между величинами, не было систематического способа определить эти взаимосвязи до того, как Фишер представил новую инновационную методологию в своей влиятельной книге *Statistical Methods for Research Workers* (1925 год)*.

* *Фишер Р.* Статистические методы для исследователей / Пер. с англ. М. : Госстатиздат, 1958. — *Прим. науч. ред.*

Анализ изменчивости в сельском хозяйстве

Фишер понял, что было необходимо провести различие между тремя типами изменчивости урожайности пшеницы: *годовой изменчивостью,* которая находилась под прямым влиянием погодных условий, стимулирующих рост растений, а также физических свойств почвы; *устойчиво-постоянной изменчивостью,* которая была связана с ухудшением питательной среды почвы; и *медленной изменчивостью,* которая была связана с небольшими непредсказуемыми изменениями.

Дисперсионный анализ и малые выборки

Дисперсионный анализ — это дисциплина и методология, связанная с набором статистических моделей для экспериментальных данных, которые подразделяют наблюдаемую изменчивость на несколько частей. Такое разделение дисперсии является ключевым для статистической методологии Фишера.

Фишер придумал **ковариационный анализ** (ANCOVA) в 1932 году для статистического контроля за переменными. Анализ заключается в том, что «ковариирует» влияние одной переменной на все другие переменные, и это может увеличить точность эксперимента, снижая дисперсию ошибки. В 1896 году Пирсон представил похожий анализ с частной корреляцией.

Статистика вывода

Основываясь на методах Пирсона, Фишер не только расширил существующую терминологию, но его статистические инновации стали основой второй фазы развития современной математической статистики благодаря его занятиям **статистикой вывода**. Если случайная вариация (или изменчивость) является основой статистики вывода, отличительной особенностью этой новой формы статистики является формальное тестирование гипотез и теория статистического оценивания.

Тестирование гипотезы — это научная процедура, позволяющая принимать рациональные решения относительно двух разных утверждений. *Теория оценивания* — это ветвь статистики, которая связана с оценкой значений параметров (см. следующую страницу), базирующейся на данных, собранных ученым. Например, политолог-аналитик хочет оценить количественное отношение совокупности голосующих в Великобритании. Это отношение является неизвестным параметром, и его оценка основана на случайной в своей основе и малой выборке голосующих.

Статистика, в которой используются латинские буквы x, s и r (для среднеарифметического, среднеквадратического отклонения и корреляции, соответственно), преимущественно была создана Пирсоном.

Параметры, обозначаемые греческими буквами μ (мю), σ (сигма малая) и ρ (ро), были введены Фишером в 1922 году для обозначения среднеарифметического, среднеквадратиччического отклонения и корреляции в генеральных совокупностях, соответственно.

Следовательно, статистика относится к выборкам, так же как параметры относятся к генеральным совокупностям.

Выборочное распределение

Для того чтобы делать обобщенные выводы о генеральной совокупности, статистическая информация берется из репрезентативной выборки.

Каждая выборка из генеральной совокупности имеет свои собственные статистические значения (X, s, или r), которые используются для оценки параметров этой генеральной совокупности (μ, σ или ρ). Согласно Фишеру, выборочная статистика должна быть несмещенной оценкой соответствующего параметра генеральной совокупности. (Фишер создал три другие оценки для параметров, которые должны были иметь свойства статистической состоятельности, эффективности и достаточности).

Для того чтобы извлечь из выборочной статистики оценку параметра генеральной совокупности, ученый использует «выборочное распределение». Вместо того чтобы пользоваться одной выборкой, несколько выборок (или даже бесконечное количество выборок) берутся из генеральной совокупности. Каждая выборка дает свои среднеарифметическое, среднеквадратическое отклонение и корреляцию. Среднее этих статистических значений по всем выборкам должно близко подходить к среднему по генеральной совокупности.

Следовательно, параметр генеральной совокупности — это способ суммирования распределения вероятностей, в то время как выборочная статистика — это способ суммирования выборки наблюдений.

Основы метода Фишера построены не только на статистических трудах Пирсона, но также являются своеобразным переводом статистического языка Пирсона. Оба они стали общеупотребительным языком современной математической статистической теории, несмотря на то, что многие из статистических методов Пирсона и его язык остаются частью собственно статистической теории.

Заключение

Бюрократическая компиляция огромного объема демографических данных викторианцами-статистиками позволила им создать статистическую систему, измеряющую здоровье нации, которая привела к политическим реформам и созданию публичных актов о здравоохранении в Британии. Идея демографических статистиков о том, что статистическая изменчивость — это дефект и источник ошибок, который нужно устранить, была оспорена идеями Чарлза Дарвина о биологической изменчивости и статистической изменчивости популяций биологических видов. Идеи Дарвина способствовали созданию новой статистической методологии, которую основал Фрэнсис Гальтон, чей интерес к измерению индивидуальных различий поставил изменчивость на передовую фронта статистики. Работы Гальтона привлекли внимание У. Ф. Р. Велдона, чьи идеи вдохновили и способствовали появлению работ Карла Пирсона и его коллег при создании основ современной математической статистики.

Первый статистический критерий контроля качества на производстве был придуман студентом Пирсона Уильямом Сили Госсетом, чьи труды вдохновили Рональда Фишера на создание статистической системы для анализа малых выборок, как следствие установив стандарты проведения статистического эксперимента и «рандомизации» в статистической теории. Развитие Фишером статистики вывода стало основой второй фазы развития современной математической статистики.

Со времен XX века статистика стала языком для медицинских, экономических и политических дискуссий. Как следствие, она проникла в повседневную речь. Статистическая информация может оказать сильное влияние на жизнь людей: на медицинское лечение, выбор машины, дома или одежды и поддержку политических партий в ходе выборов. В движимом технологиями информационном XXI веке понимание статистики остается первостепенным элементом жизни.

Глоссарий основных статистических терминов и понятий*

Agricultural variation — изменчивость в сельском хозяйстве (в т. ч. в земледелии)
Analysis of covariance (ANCOVA) — ковариационный анализ
Analysis of variance (ANOVA) — дисперсионный анализ
Association — ассоциация, ассоциативная связь
Association factor — коэффициент ассоциации (связи переменных)
Averages — средние значения (в противопоставление «изменчивости», variation)

Bayesian approach — Байесовский подход к анализу вероятностей
Bimodal distribution — бимодальное распределение
Binomial distribution — биномиальное распределение
Biserial correlation — бисериальная корреляция

Categories — категории
Causation — причинность
Central limit theorem — центральная предельная теорема
Chi-square system — система хи-квадрат
Clustered data — данные, распределенные по группам; сгруппированные по кластерам
Coefficient of variation — коэффициент изменчивости, коэффициент вариации
Contingency tables — таблицы сопряженности (признаков)
Continuous / discrete data (variables) — непрерывные / дискретные данные (переменные)
Continuous / discrete distribution — непрерывное / дискретное распределение
Correction factor — поправочный коэффициент (Пирсона)
Correlation — корреляция
Correlation ratio — корреляционное соотношение (Пирсона)
Covariance — ковариация
Curve-fitting for asymmetrical distributions — подгонка (подбор) кривой для асимметричных распределений
Curvilinear relationships — взаимосвязи переменных, представленные разными кривыми

Data management procedures — статистические процедуры управления данными
Degrees of freedom — степени свободы
Demography — демография
Dependent variables — зависимые переменные
Determinism — детерминизм (в статистике)
Developmental correlation — корреляция развития (в эволюционной биологии)
Dichotomies — дихотомии
Discrete data (variables) — дискретные данные (переменные)
Directional selection — направленный отбор, направленная селекция
Disruptive selection — разрывающий отбор, разрушающая селекция
Distribution — распределение (вероятностей)

Ecological correlations — экологические корреляции (в эволюционной биологии)
Error curve — кривая ошибок
Estimation theory — теория статистического оценивания
Expected values — ожидаемые значения

F-distribution — F-распределение
Factor analysis — факторный анализ
Frequency distribution — плотность вероятности, распределение частот (частостей)
Frequency polygon — полигон частот
Functional correlations — корреляции, выраженные в виде функций

Games of chance — игры случая
Gaming theory — азартная игра, пари
Gaussian curve — кривая Гаусса (кривая нормального распределения)
Gaussian distribution — распределение Гаусса (нормальное распределение)
General Register Office — Управление записи актов гражданского состояния (в Великобритании)
Goodness of fit test — статистический критерий согласия

Histogram — гистограмма
Hypothesis testing — тестирование, проверка статистических гипотез

Incidental sampling — побочная выборка
Independent / dependent variables — независимые / зависимые переменные
Inferential statistics — статистика (логического) вывода
Insurance statistics — страховая статистика
Interquartile range — интерквартильный размах

Kendall's tau — тау-коэффициент Кендалла (непараметрический коэффициент ранговой корреляции)
Kruskall-Wallis analysis of ranks — анализ рангов Крускала=Уоллиса
Kurtosis — эксцесс, коэффициент эксцесса

Least squares method — метод наименьших квадратов
Lexican ratio, L — соотношение Лексиса

Malthusian populations — Мальтузианское население
Mann-Whitney U test — U-критерий Манна-Уитни
Mathematical statistics — математическая статистика
Matrix algebra — матричная алгебра
Mean, arithmetical mean — арифметическое среднее, среднеарифметическое
Median — медиана, медианное значение
Mesokurtic curves — кривые распределения с нулевым или нормальным эксцессом
Method of moments — метод моментов (Пирсона)
Mode — мода, модальное значение
Mortality statistics — статистика смертности
Mortality tables — таблицы смертности

* Составлен научным редактором данного издания доктором экономических наук, профессором НИУ ВШЭ и Финансового университета П.Н. Клюкиным.

Multiple correlation / regression — множественная корреляция / регрессия

Natural selection — естественный отбор (в учении Ч. Дарвина)
Negative correlation — отрицательная корреляция
Nominal / ordinal variables — номинальные / порядковые переменные
Normal distribution — нормальное распределение
Normal curve — кривая нормального распределения

Odds ratio — коэффициент несогласия (основан на Q-статистике Юла)
Ordinal / nominal variables — порядковые / номинальные переменные
Outliers — выпадающие значения, статистические выбросы

Part correlation — частная корреляция
Path analysis — пат-анализ
Pearson product-moment correlation coefficient — коэффициент корреляции Пирсона
Pearsonian family of curves — семейство кривых Пирсона
Percentile — процентиль
Philosophy of statistics — философия статистики
Platykurtic curves — кривые распределения с отрицательным эксцессом
Plotted frequency diagram — график плотности распределения, диаграмма частот
Point-biserial correlation — точечно-бисериальная корреляция
Poisson distribution — распределение Пуассона
Polychoric correlation — полихорическая корреляция
Population — генеральная совокупность
Population distribution — распределение генеральной совокупности
Principles of least squares — принцип наименьших квадратов
Probability — вероятность
Probability distribution — распределение вероятностей
Probability mass function — функция распределения масс (распределение вероятностей дискретной случайной величины)
Probability tables — таблицы значений вероятности
Product-moment correlation — корреляция произведения моментов (К. Пирсона)
Purposive sampling — целевая выборка
Quantities — количества
Quartile — квартиль
Quetelismus — Кетлесимус (эпоха господства идей Кетле)

Random sampling — случайная выборка
Random variables — случайные переменные
Randomization — рандомизация, метод случайного отбора
Range — диапазон, размах
Range deviation — отклонение на интервале
Rank order correlation — корреляция рангов, ранговая корреляция
Regression coefficient — коэффициент регрессии
Regression line — линия регрессии
Relative / absolute measure of variation — относительная / абсолютная мера (характеристика) изменчивости
Relative frequency — относительная частота (частость)

Saltational origins — происхождение (видов) в рамках теории скачкообразной динамики
Sample — выборка
Sample size — размер выборки
Sample statistics — выборочная статистика
Sampling distribution — выборочное распределение
Sanitary Reforms — санитарные реформы
Scales of measurement — шкалы измерения, измерительные шкалы
Scatter diagrams — диаграммы рассеяния
Semi-interquartile range — полуинтерквартильный размах
Significance testing — критерий статистической значимости
Simple / multiple correlation — простая / множественная корреляция
Skewed distributions — асимметричные распределения
Skewness — асимметрия, коэффициент асимметрии
Small / large samples — малые / большие выборки
Spearman rho — коэффициент «ро» Спирмена
Species — вид (в теории Дарвина)
Spurious correlation — ложная (кажущаяся) корреляция
Stabilizing selection — стабилизирующий отбор
Stable / unstable conditions — устойчивые / неустойчивые состояния, стабильные / нестабильные режимы
Standard deviation — среднеквадратическое отклонение, стандартное отклонение
Standardized frequency distributions — нормированное распределение частот (частостей), плотность нормированного распределения, унифицированное частотное распределение
Statistical distribution — статистическое распределение
Statistical variation — статистическая изменчивость
Statistical control — статистический контроль
Statistical data — статистическое данные
Statistical measures of variation — статистические характеристики (меры) изменчивости
Statistical quality control tests — статистический критерий контроля качества
Stratified sampling — расслоенная (типологическая) выборка
Student's t-distribution — t-распределение Стьюдента (Госсета)
Subjective approach — субъективный подход (к вероятности)
Systematic sampling — систематическая выборка

t-distribution — t-распределение (Стьюдента)
Tetrachoric correlation coefficient — тетрахорический коэффициент корреляции
Theory of errors — теория ошибок
Triserial correlation — трисериальная (трехрядная) корреляция

Variables — переменные
Variance — дисперсия
Variation — изменчивость, вариация
Vital statistics — демографическая статистика

Wilcoxon signed-rank test — критерий знаковых рангов Уилкоксона

Yule's Q-statistic — Q-статистика Дж. У. Юла

Z-ratio (test) — z-соотношение Госсета (т. е. Стьюдента)

Издание для дополнительного образования

БИЗНЕС В КОМИКСАХ

Магнелло Эйлин, Ван Лоон Борин

СТАТИСТИКА В КОМИКСАХ

Руководитель отдела *О. Усольцева.* Ответственный редактор *Л. Ивахненко*
Выпускающий редактор *К. Ананьева.* Научный редактор *П. Клюкин*
Художественный редактор *В. Брагина.* Технический редактор *М. Печковская*
Компьютерная верстка *С. Пяташ.* Корректор *М. Козлова*

ООО «Издательство «Эксмо»
123308, Москва, ул. Зорге, д. 1. Тел.: 8 (495) 411-68-86.
Home page: www.eksmo.ru E-mail: info@eksmo.ru
Өндіруші: «ЭКСМО» АҚБ Баспасы, 123308, Мәскеу, Ресей, Зорге көшесі, 1 үй.
Тел.: 8 (495) 411-68-86.
Home page: www.eksmo.ru E-mail: info@eksmo.ru.
Тауар белгісі: «Эксмо»
Интернет-магазин : www.book24.kz
Интернет-дүкен : www.book24.kz
Импортёр в Республику Казахстан ТОО «РДЦ-Алматы».
Қазақстан Республикасындағы импорттаушы «РДЦ-Алматы» ЖШС.
Дистрибьютор и представитель по приему претензий на продукцию,
в Республике Казахстан: ТОО «РДЦ-Алматы»
Қазақстан Республикасында дистрибьютор және өнім бойынша арыз-талаптарды
қабылдаушының өкілі «РДЦ-Алматы» ЖШС,
Алматы қ., Домбровский көш., 3«а», литер Б, офис 1.
Тел.: 8 (727) 251-59-90/91/92; E-mail: RDC-Almaty@eksmo.kz
Өнімнің жарамдылық мерзімі шектелмеген.
Сертификация туралы ақпарат сайтта: www.eksmo.ru/certification

Сведения о подтверждении соответствия издания согласно законодательству РФ
о техническом регулировании можно получить на сайте Издательства «Эксмо»
www.eksmo.ru/certification
Өндірген мемлекет: Ресей. Сертификация қарастырылмаған

Подписано в печать 24.05.2018.
Формат 70×100 1/16. Гарнитура «SansRoundedLight».
Печать офсетная. Усл. печ. л. 14,26.
Тираж экз. Заказ

ISBN 978-5-04-090149-4

КОГДА ВЫ ДАРИТЕ КНИГУ, ВЫ ДАРИТЕ ЦЕЛЫЙ МИР

ХОТИТЕ ЗНАТЬ БОЛЬШЕ?

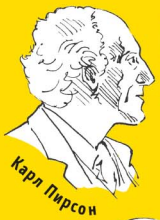
Заходите на сайт:
https://eksmo.ru/b2b/

Звоните по телефону:
+7 495 411-68-59, доб. 2261

ВАШ ЛОГОТИП НА ОБЛОЖКЕ

ВАШ ЛОГОТИП НА КОРЕШКЕ

ОБРАЩЕНИЕ К КЛИЕНТАМ НА ОБЛОЖКЕ



Карл Пирсон

В этом комиксе есть что-то про выборки?

Да!

Чарльз Дарвин

А говорится, что мои идеи легли в основу методов статистики?

Да...

Марк Твен

А моя знаменитая цитата?

Конечно!

В этом комиксе основные концепции статистики, обзор истории науки и то, как она связана с реальными проблемами. Медианные значения, распределения, пат-анализ, дисперсия, «жизненная статистика» Уильяма Фарра и математическая Карла Пирсона... – основные понятия и теории с отличным юмором и иллюстрациями.

Решения, основанные на статистике, принимаются каждый день и влияют на нашу повседневную жизнь. От тестов на профпригодность, которые дают нам работодатели, одежды, которую мы носим, до машин, которые мы водим, и даже пива, которое мы пьем. Знание основ статистики может даже спасти или продлить жизни!

MUST-READ ДЛЯ КАЖДОГО, ЧЬЯ ЖИЗНЬ, УЧЕБА ИЛИ РАБОТА ХОТЬ НЕМНОГО СВЯЗАНЫ С ЧИСЛАМИ

ISBN 978-5-04-090149-4

БОМБОРА

Бомбора — это новое название Эксмо Non-fiction, лидера на рынке полезных и вдохновляющих книг. Мы любим книги и создаем их, чтобы вы могли творить, открывать мир, пробовать новое, расти. Быть счастливыми. Быть на волне.

bomborabooks
www.bombora.ru

БИЗНЕС В КОМИКСАХ ДЛЯ ГИКОВ